Structural studies on protein targets from the pathogenic bacterium *Burkholderia pseudomallei*

Matthew Day

B.Sc. (Hons) Biochemistry, University of Sheffield

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Molecular Biology and Biotechnology,

University of Sheffield

September 2012

Abstract

The main body of this thesis deals with a small-scale structural genomics project for proteins from the bacterium, *Burkholderia pseudomallei*. Potential pathogenicity determinants were selected based on their possession of at least one of two criteria. The first was their expression in the pathogenic *Burkholderia pseudomallei* but not the closely related, non-pathogenic *Burkholderia thailandensis*. The second required their expression be controlled as part of a stress response, which controls other known virulence factors. Of the nine selected targets, five were successfully overexpressed, purified and crystallised and the structure of one was determined. The selected targets were all of unknown function and initial sequence analysis highlighted a number of interesting characteristics for certain targets.

BPSL1958 contains a highly repetitive sequence conserved at both the protein and DNA level, suggesting the gene was created in a recent duplication event and there has not been enough time for the sequence of the individual repeats to drift.

Four of the selected targets, BPSS0211 – BPSS0214, formed an operon, of these BPSS0212 and BPSS0213 were homologous, with BPSS0211 also representing the C-terminal domain of the two proteins. The structure of BPSS0211 was determined to 2.17 Å using multiple-wavelength anomalous dispersion experimental phasing. BPSS0211 is a small protein annotated as containing a conserved domain of unknown function, DUF1843. The tertiary structure of BPSS0211 entails two alpha helices stacked together which assemble to form a tetramer comprising a dimer of dimers. The pattern of conservations between residues in DUF1843 within BPSS0211, BPSS0212 and BPSS0213 is such that interactions between the three proteins would be possible. The proposed function of DUF1843 is as an oligomerisation domain, allowing the formation of homo and hetero, dimer and tetramer, complexes of three proteins encoded within the BPSS0211-BPSS0214 operon.

Other work included involves the determination of the structure of thioredoxin from *Burkholderia pseudomallei* as part of an on-going project into this system. The structure was solved to 1.07 Å using molecular replacement and agreed well with previously solved thioredoxin structures from other organisms.

Acknowledgements

Firstly I would like to thank my supervisor, Professor David Rice, for his support and guidance during my PhD. I am also indebted to many other members of the Sheffield crystallography group, for their help and advice throughout the course of my PhD studies. I would particularly like to thank Dr Svetlana Sedelnikova for her protein purification expertise and Dr Patrick Baker for his tireless manipulation of crystals that made the results obtained possible. I should also mention Professor Peter Artymiuk and Dr John Rafferty for advice provided whenever requested and Fiona Rodgers, without whose organisation the lab would fall apart. I am also grateful to Dr John Rafferty for overseeing my undergraduate project, the experience of which led me to pursue a PhD. I would also like to thank all the other members of the group, past and present, who have helped me along the way and thanks also go to the BBSRC for funding this project.

Finally I must acknowledge Jose Cree, for her love and support, throughout the many years we were together in Sheffield.

Declaration

The work described in this thesis was carried out in the Department of Molecular Biology and Biotechnology at the University of Sheffield from October 2008 to September 2012. It is the original work of the author, unless otherwise acknowledged within the text. It has not been submitted previously for a degree at this or any other university.

Matthew Day

September 2012

Table of Contents

Chapter one: Introduction to the bacterial pathogen, *Burkholderia pseudomallei*, and a small-scale structural genomics project conducted as part of this thesis.

List of figures

List of tables

Chapter one

Introduction to the bacterial pathogen, *Burkholderia pseudomallei*, and a small-scale structural genomics project conducted as part of this thesis

Section 1 *Burkholderia pseudomallei* is the causative agent of melioidosis

Section 2 Molecular pathogenicity determinants of *Burkholderia pseudomallei*

Section 3 Project aims, target selection and sequence analysis

1.0 *Burkholderia pseudomallei* is the causative agent of melioidosis

This section provides an introduction about the bacterial pathogen *Burkholderia pseudomallei* and its biological relevance as the causative agent of the human disease melioidosis. Melioidosis is defined as any disease caused by infection with the bacterium *Burkholderia pseudomallei* and was first reported as the cause of death in the post-mortem of a morphine addict in Rangoon, Burma in 1911. Detailed microbiological studies identified a new organism, *Burkholderia pseudomallei*, which fulfilled Koch's postulates as the causative agent of the disease [1, 2].

1.1 Taxonomy

The *Burkholderia* genus was created recently in 1992 with the transfer of seven species from the *Pseudomonas* genus [3]. It now consists of over sixty species occupying a large range of ecological niches [4] with highly diverse genomes [5]. The evolutionary history of the genus is complex, with high levels of lateral gene acquisition and transfer between species inside the genus, broader bacteria and some eukaryotic organisms [6]. The phylogeny has been mapped using various techniques [7, 8] one of which is the sequence divergence of the *recA* gene in the individual species [9] (figure 1.1). The genus consists of two genetic lineages with one representing mainly pathogenic bacteria and the other predominantly non-pathogenic environmental isolates often symbiotically associated with plant species. The pathogenic group can be split into three further sub-groups. The *Burkholderia cepacia* complex contains a number of opportunistic pathogens that can be responsible for lung infections in cystic fibrosis sufferers. The phyto-pathogenic sub-group contains a number of plant pathogens, some of which are responsible for disease in important food crops. The last sub-group, the *pseudomallei* group, is the most important in terms of human disease and contains the three species, *Burkholderia pseudomallei*, *mallei* and *thailandensis*.

1.1.1 The pseudomallei group

Burkholderia pseudomallei and *mallei* are the causative agents of the human disease melioidosis and the equine disease glanders respectively. *Burkholderia mallei* is unable to persist outside of a host while *pseudomallei* and *thailandensis* occupy the same environmental soil niche in endemic areas. *Burkholderia thailandensis* is a non-pathogenic species, being more than 10,000 fold less virulent than *pseudomallei* in model organisms [10]. The three closely related species are thought to have recently evolved from a common

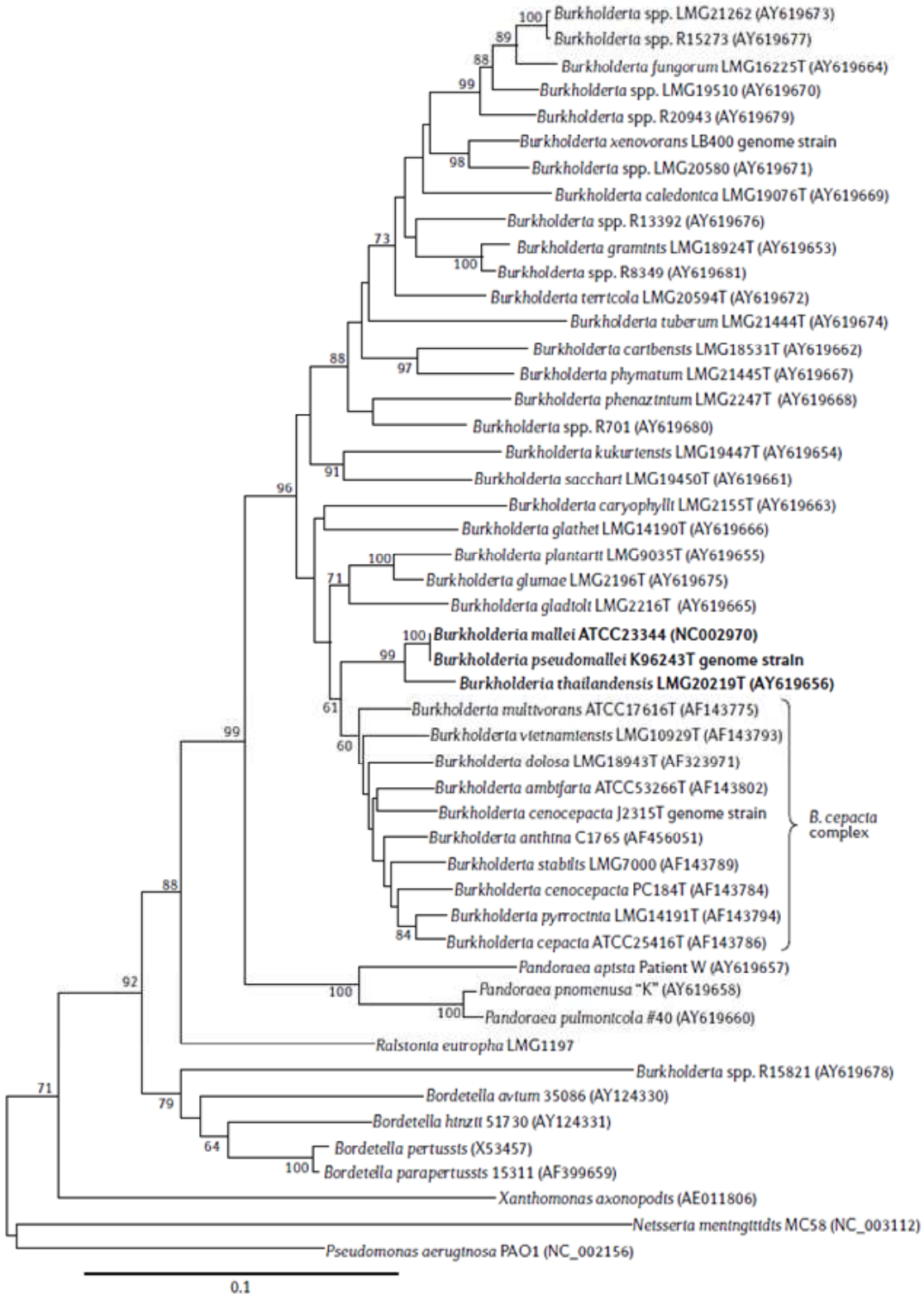


Figure 1.1 Phylogenetic tree for the *Burkholderia* genus based on *recA* sequences. The tree is rooted using the recA gene from *Pseudomonas aeruginosa* PAO1 and includes 27 *Burkholderia* species (those in bold represent the pseudomallei group) and 11 other closely related organisms. Bootstrap values and genetic distance are shown. Figure adapted from "Development of a *recA* gene-based identification approach for the entire *Burkholderia* genus" [9].

Australian ancestor followed by an introduction into South East Asia and further divergence into numerous strains [11, 12]. Understanding the difference between these closely related species and strains would answer important questions, providing a window into the differing level of pathogenicity, host specificity, virulence determinants and the ability of the bacteria to persist in the environment.

1.2 Bacteriology

Burkholderia pseudomallei are gram-negative, non-spore-forming, aerobic, motile, rod-shaped bacteria (figure 1.2). The bacteria exist in the environment as a soil-dwelling saprophyte as well as a pathogen of a number of organisms. Melioidosis has been identified in a wide range of animals including mammals, birds, fish and reptiles with varying levels of severity [13]. *Burkholderia pseudomallei* are also capable of infecting certain plant species including tomatoes providing a further environmental reservoir [14]. The organism is highly resilient in the environment with the ability to survive in a wide range of harsh conditions in both soil and liquid media [15]. The bacteria can not only survive but continue to grow under a diverse range of temperature [16], acid and alkali pH [17], and severe dehydration conditions [16]. *Burkholderia pseudomallei* are also able to survive extreme nutrient deprivation [18] with the bacteria still maintaining viability after at least sixteen years in distilled water [19]. *Burkholderia pseudomallei* can survive exposure to disinfectant and antiseptic solutions [20, 21] and exhibit a tolerance to chlorinated water with levels used in water supplies proving only bacteriostatic [22] although elevated levels are effective at killing the bacteria [23]. *Burkholderia pseudomallei* are also naturally resistant to a large number of antibiotics including cephalosporins, penecillins, rifamycins, aminoglycosides, quinolones and macrolides [15].

1.2.1 Colony morphology

Clinical and environmental isolates of the bacteria exhibit a diverse morphology with seven colony types (figure 1.3) when plated out on Ashdown's agar [24]. Morphology is not strain dependant with switching between types being inducible under certain conditions and reversible within the same strain. Different morphotypes have different proteomic profiles [25] and levels of virulence [26]. Environmental strains almost always display colony morphology I and there is a relationship between isolate morphology and patient history. Morphology I is the most commonly identified in samples from initial infections, with morphologies II and III being represented more often from relapse patients [26]. The level of

phenotypic change may provide a mechanism for the bacteria to persist in the host in a latent infection or survive antibiotic treatment and subsequently cause relapse in patients.

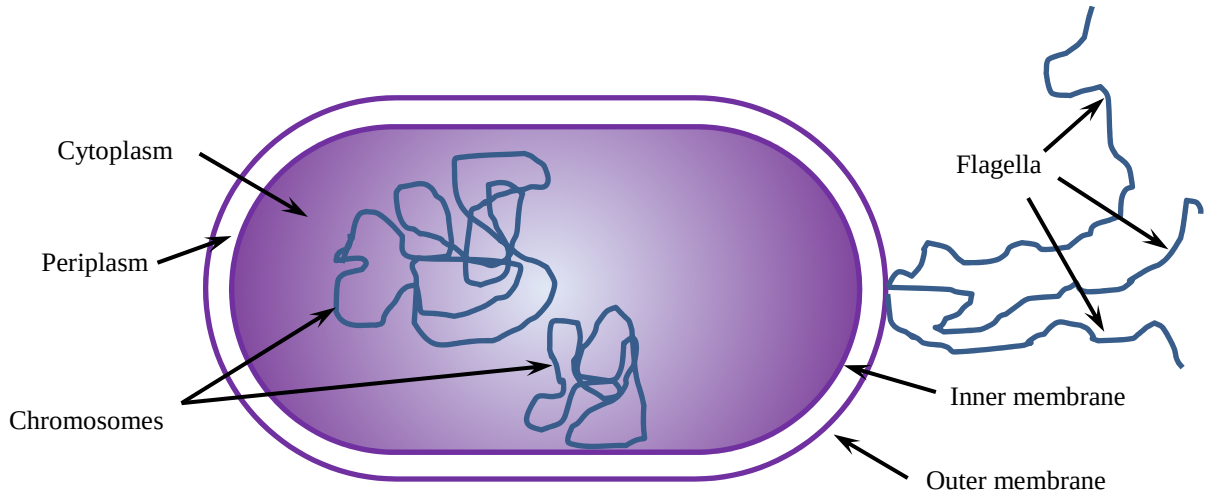


Figure 1.2 Schematic diagram showing the cellular structure of *Burkholderia pseudomallei*.

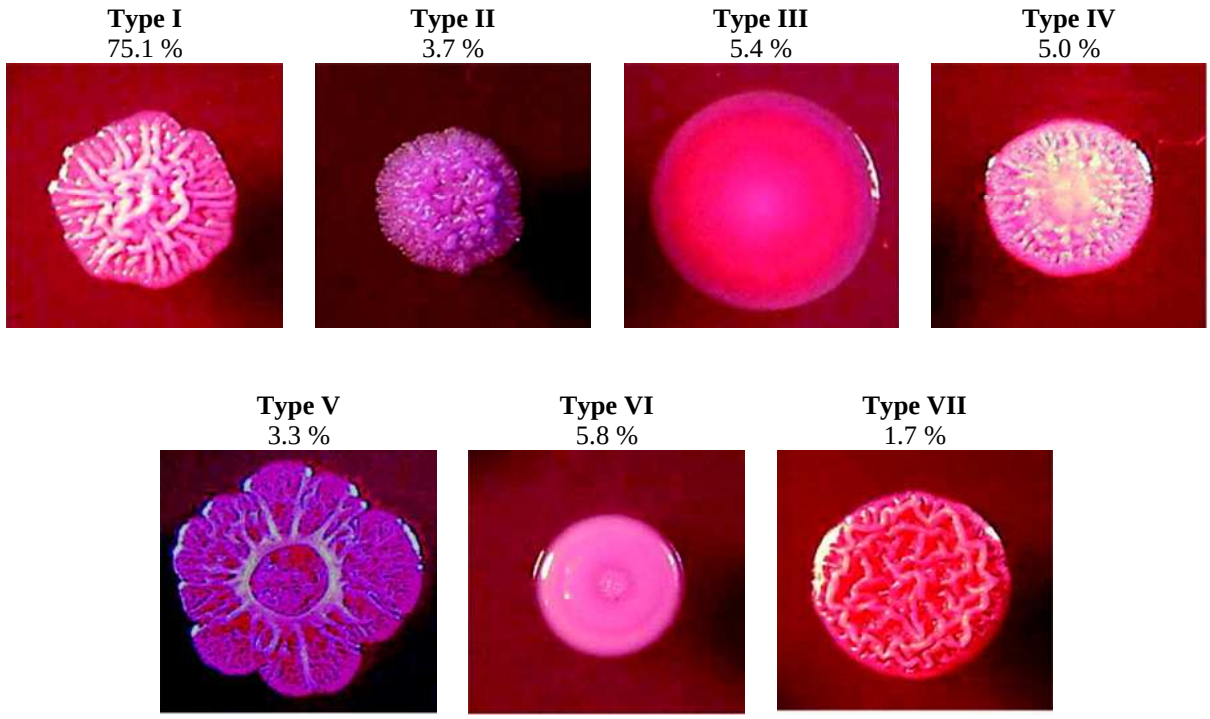


Figure 1.3 The seven different colony morphologies of *Burkholderia pseudomallei*. Colonies were morphotyped based on appearance, size and pigmentation. The percentages indicate the proportion of each form in a collection of 212 clinical isolates after growth for four days on Ashdown's agar. Figure adapted from "Biological relevance of colony morphology and phenotypic switching by *Burkholderia pseudomallei*" [24].

1.3 Epidemiology

Melioidosis is endemic throughout the tropical regions of South East Asia and Northern Australia where *Burkholderia pseudomallei* is found in the environment [15] (figure 1.4). It is thought that the extent of the disease in these areas as well as worldwide is misrepresented, with many cases going unrecognised due to lack of awareness and diagnostic facilities in affected areas, particularly amongst rural communities [27]. There is evidence for emerging areas of endemicity around the globe, probably due to wider awareness of the disease in many areas particularly tropical regions of Africa, the Americas, and a growing area of South East Asia and the wider South Pacific [28, 29].

1.3.1 Cellular basis of disease

Burkholderia pseudomallei is an intracellular pathogen with the ability to invade, survive and multiply inside host phagocytic and non-phagocytic cells [30] (figure 1.5). Before the bacteria can enter into a host epithelial cell it must first interact with the cells surface through adhesion. This is mediated by flagella based motility, the bacterial capsule and through the use of type IV pili. The bacteria are taken up into vacuoles, by phagocytosis or invasion, which they can later escape from into the host cells cytoplasm. Once inside host cells *Burkholderia pseudomallei* are able to survive, avoiding the bacteriocidal activities of immune cells and the induction of autophagy. The bacteria utilise a form of ARP2/3 independent actin based motility to travel inside host cells [31]. This form of motility also allows the bacteria to travel between cells and is involved in the formation of multinucleated giant cells. *Burkholderia pseudomallei* is capable of inducing cell lysis through apoptosis or caspase-1-dependent cell lysis [32]. This releases the bacteria from the cell allowing invasion of further host cells. The bacteria are able to spread to other areas of the body by transport either in the lymphatic or circulatory systems.

1.3.2 Modes of acquisition

Infection with *Burkholderia pseudomallei* usually occurs through cutaneous inoculation or inhalation of infected soil or water. The disease can be obtained through ingestion [33] and it has also been rarely recorded as transmitting through human to human contact, either sexually or during childbirth [34, 35]. Outbreaks of melioidosis are known to exhibit seasonal variability with an increased frequency during the rainy season due to increased contact with the bacterium through agricultural activity and extreme weather events [30]. Several clinical risk factors are known to provide a predisposition to developing melioidosis particularly

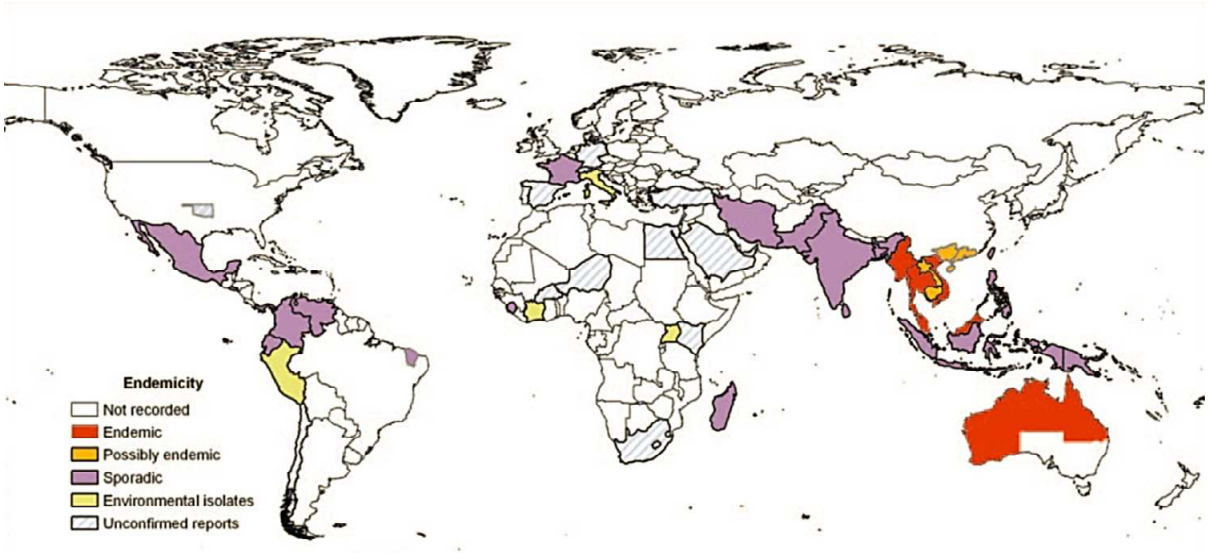


Figure 1.4 Map representing a summary of known global melioidosis endemicity in 2005. Figure adapted from "Melioidosis: Epidemiology, pathophysiology, and management" [15].

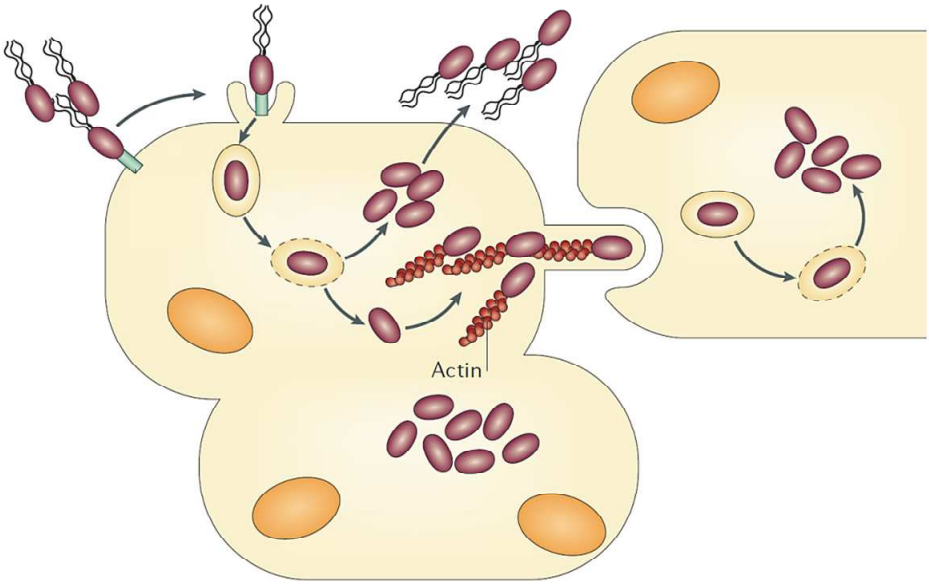


Figure 1.5 The intracellular lifestyle of *Burkholderia pseudomallei*. Bacteria enter into host cells inside primary phagosomes by invasion or phagocytosis. During phagosomal maturation bacteria secrete effectors that lead to the disruption of vacuolar membranes and escape into the host cells cytosol. Here the bacteria can multiply, induce cell lysis, engage actin based motility and spread to other cells by inducing cell fusion or pushing through the cell membrane and entering neighbouring cells in a secondary vacuole. Figure adapted from "Melioidosis: insights into the pathogenicity of *Burkholderia pseudomallei*" [30].

diabetes mellitus, thalassemia and renal disease [15]. In rural rice farming communities in Thailand, the majority of the population are seropositive for *Burkholderia pseudomallei* [36]. However seropositivity does not provide protection against further infection, and it is unclear what proportion of people testing positive have cleared their systems of the bacteria or are harbouring a latent infection.

1.3.3 Clinical manifestation

Melioidosis can present as a wide range of disease states from acute septicaemia to a chromic localised infection which has earned it the nickname 'the great mimicker' [37]. The disease can affect almost any organ or tissue resulting in an acute or chronic disease state. Pneumonia represents the most common clinical manifestation of melioidosis either resulting from initial infection via inhalation of the bacterium or following septiceamic spread from another infected site. Chronic lung infection is also common presenting a tuberculosis type disease. Other possible sites of infection include the liver, spleen, genito-urinary tract (leading to prostatic abscess), skin, soft tissues, bones and skeletal muscle. Disease manifestation varies throughout the world possibly due to different strains of *Burkholderia pseudomallei* existing in the environment. There is a high incidence of acute suppurative parotiditis amongst paediatric cases in Thailand, with the condition being almost absent in Australia, while encephalomyelitis and genito-urinary infections are common in Australia but uncommon throughout South East Asia. The disease also can exist in an asymptomatic state for long periods of time with clinical manifestations often occurring decades after initial infection, usually triggered by a weakening of the host defence by other diseases.

1.3.4 Current treatment

Current best practice treatment for melioidosis lasts twenty weeks and is divided into two stages, an initial high intensity intravenous phase of treatment followed by a prolonged oral eradication phase. Initial treatment varies by region but at its core involves a high dose of intravenous ceftazidime for at least ten days or until improvement occurs. It is not uncommon for the intravenous stage of treatment to proceed for over a month before the fever has subsided for more than two days and the eradication phase can begin. Without treatment acute melioidosis has a mortality rate exceeding 80 %, the initial treatment alone reduces overall mortality by more than 50 % [38]. The eradication phase also varies by region but typically consists of a two drug combination of trimehoprim and co-trimoxazole with the possible addition of chloramphenicol and doxycycline for the initial eight weeks of the

treatment [15, 39]. Relapse into the disease state is common amongst patients and is caused by the reactivation of the original infection following failure to clear the infection. The major factors that contribute to relapse are the initial severity of the disease and a failure to adhere to the full course of antibiotic treatment, due to the prolonged course, cost and adverse side effects associated with the treatment [15]. A full course of antibiotic treatment results in a relapse rate of approximately 10 % with this increasing to 30 % in patients who fail to complete at least 8 weeks of eradication phase antibiotic treatment [39].

1.3.5 New treatments are required

The current treatment for melioidosis is far from ideal with the mortality rate remaining high, the prolonged course and cost of treatment, and the emergence of a number of ceftazidime resistant strains [40-42]. The need for a better understanding of the organism and the creation of new treatments is clear. There is a possibility that bactriophage could be utilised as an effective treatment in the future with the identification of a specific podovirus capable of lysing several clinically relevant strains of *Burkholderia pseudomallei* but not other species of the *Burkholderia* genus [43].

1.3.6 Vaccine development

Several approaches are being undertaken in order to develop effective vaccines for use against melioidosis [44, 45]. Several live attenuated mutants have been used as vaccines in mice models providing high levels of resistance. However these are unlikely to be developed for human use due to fear of reversion and the bacteria's ability to establish a persistent latent infection which can remain dormant for prolonged periods of time. Inactivated whole cell vaccines studies in mice using either *Burkholderia pseudomallei* or the non-pathogenic *thailandensis* have resulted in protective immunity, however as with all inactivated whole cell vaccines there is potential for undesirable side effects. A number of individual protein targets have been tested for their ability to induce immunity in mice models with some success. The use of naturally derived outer-membrane vesicles has also shown promise as a potential vaccine [46]. The use of *Burkholderia pseudomallei* DNA as a potential vaccine has been tested successfully in mice though potential DNA vaccines against other organisms that showed promise in mice were found to be ineffective in human trials. The use of monoclonal antibodies specific to the bacteria lipopolysaccharide and capsular polysaccharide for passive immunisation in mice has been shown to offer protection from infection with the bacterium [47]. Overall the hope of finding a long lasting, totally effective vaccine against *Burkholderia*

pseudomallei seems unlikely as repeated exposure to environmental isolates of the bacterium in people inhabiting the endemic areas does not prevent further infection [39]. However, even if a vaccine was not totally effective at preventing disease incidence it could represent a cost effective public health initiative particularly if administered to at risk groups due to the high cost of treatment [45].

1.4 *Burkholderia pseudomallei* as a potential bioweapon

Burkholderia pseudomallei's ability to survive a large range of conditions, it's high mortality rate, difficulty to treat and lack of an effective vaccine has led to it being considered as a potential bioweapon. The United States Centre for Disease Control has classified *Burkholderia pseudomallei* as a category B potential biological threat. This places the organism in the second highest priority category alongside other potentially air or water borne pathogens and threats to global food and water safety [48].

1.5 Genomics

The first reported full genome sequence was for the clinical isolate strain K96243 and comprised two chromosomes of 4.07 and 3.17 megabase pairs (figure 1.6 a) [49]. The gene nomenclature used in this study is based on this genome sequence with the 3,460 genes on the larger chromosome being designated BPSL____ and the 2,395 genes on the smaller chromosome BPSS____. The two chromosomes were found to contain genes broadly involved in different roles, with the larger chromosome encoding many of the core functions, and the smaller chromosome genes involved in adaptation and survival in different niches, although genes from all functional classes were found on both chromosomes (figure 1.6 b). The genome contains 16 identifiable genomic islands making up 6.1 % of the total genome with properties suggesting recent lateral gene acquisition by *Burkholderia pseudomallei*. A statistical analysis of the K96243 evolutionary history found very high levels of lateral gene acquisition with 109 genomic islands of different predicted ages highlighting the importance of horizontal gene transfer in the evolution of this species [6]. The genome contains a large proportion of genes of unknown function, accounting for over 20 % of annotated genes within the genome, demonstrating the need for studies into the basic biology of *Burkholderia pseudomallei*.

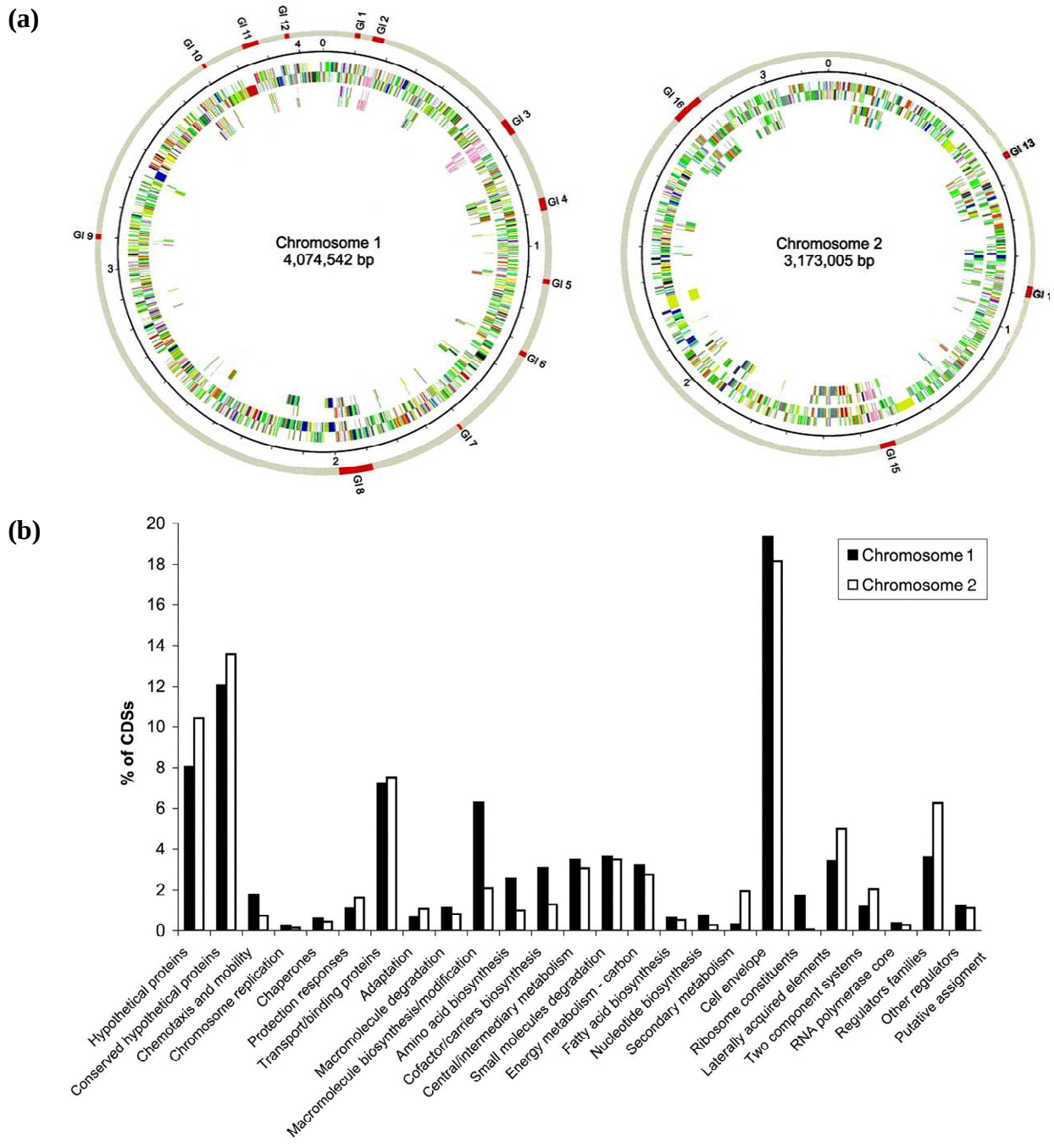


Figure 1.6 Genome sequence of *Burkholderia pseudomallei* strain K96243. **a** Schematic representation of the large and small chromosomes with the genomic islands shown as red segments. Genes are colour coded by predicted function: dark blue, pathogenicity/adaptation; black, energy metabolism; red, information transfer; dark green, surface-associated; cyan, degradation of large molecules; magenta, degradation of small molecules; yellow, central/intermediary metabolism; pale green, unknown; pale blue, regulators; orange, conserved hypothetical; brown, pseudogenes; pink, phage plus IS elements; grey, miscellaneous. **b** Distribution of genes belonging to different functional classes within the *Burkholderia pseudomallei* genome sequence expressed as a percentage of the total genes on each chromosome. Figure adapted from "Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*" [47].

2.0 Molecular pathogenicity determinants of *Burkholderia pseudomallei*

The *Burkholderia pseudomallei* genome contains a multitude of genes that encode factors involved in the bacterium's pathogenicity or survival. This section details those which have been annotated in the genome and/or experimentally characterised.

2.1 Quorum sensing

Cell density dependent cell to cell communication systems allow bacteria to co-ordinate gene expression through the use of small signalling molecules. One system used by many species of bacteria involves the use of N-acyl-homoserine lactones (AHLs). *Burkholderia pseudomallei* contains three LuxI genes, responsible for AHL synthesis, and five LuxR genes, which act as transcriptional regulators. These systems control the expression of a number of genes including metalloproteases, phospholipase C and DspA. The system is also important for the production of biofilms [50]. Mutants in any of the LuxI or LuxR genes are less virulent with increased survival times of infected mice and an increased lethal dose 50 % in hamster models [51]. There is evidence for the requirement of the Bpe-OprB multidrug efflux pump for secretion of AHLs from inside the cell [52] [53]. A second quorum sensing system has also been identified in *Burkholderia pseudomallei* involving the production, release and detection of 4-hydroxy-3-methyl-2-alkylquinolones (HMAQs) [54]. The function this second system plays in regulation and virulence is yet to be determined.

2.2 Alternate sigma factors

The multisubunit RNA polymerase of eubacteria is required to associate with a sigma factor in order to initiate transcription. The sigma factor is responsible for promoting recognition of the transcription start site allowing translation to start. Alternative sigma factors control the expression of many genes associated with response to a variety of stresses and of virulence factors in some pathogenic bacteria. *Burkholderia pseudomallei* contains seventeen sigma factors in its genome, two of which, RpoE (BPSL2434) and RpoS (BPSL1505) have begun to be characterised. RpoE is required for response to several environmental stresses including heat, osmotic shock and H_2O_2 [55-57]. The response regulator also has a role in the production of biofilms [55] and has been linked to pathogenesis by some genes it controls [57] and as knockout mutants have reduced survival in macrophage cell lines [55]. The transcription of RpoS is controlled by growth phase with an up regulation upon entering stationary phase [58]. RpoS is required for resistance to carbon starvation, acidic conditions

and oxidative stress but not heat or osmotic shock [58, 59]. It has been directly linked to pathogenicity as it is required for cellular invasion, the formation of multinucleated giant cells and inhibition of inducible nitric oxide synthase expression within macrophage [60]. The system is also required for the induction of host cell apoptosis [61]. RpoS also plays a role in regulating the production of AHLs in quorum sensing and several genes are under the control of both systems [62].

2.3 Capsule

Bacterial capsules consist of an organised layer, usually consisting of polysaccharides, arranged outside the cell wall. There are three morphologically distinct forms of *Burkholderia pseudomallei* based on its polysaccharide capsule [63]. One form displays no capsule with two others having capsules of differing thickness. The thicker capsular form (0.10 – 0.25 µm) most likely represents the initial stage of biofilm formation with the thinner morphology (0.09 µm) the usual unicellular form. Four separate capsular polysaccharide structures (CPS) have been observed [64-68] and their biosynthetic clusters have been identified in the genome [69] (BPSL2786 – BPSL2810, BPSS0417 – BPSS0429, BPSS1825 – BPSS1835 and BPSL2769 – BPSL2785). CPS I is found in *Burkholderia pseudomallei* and *mallei* with CPS II – IV found in *pseudomallei* and *thalandensis*. The distribution of capsule biosynthetic clusters in the three species suggests CPS I may be involved in pathogenicity with CPS II – IV playing a role in survival in the environment. CPS I protects against C3b complement deposition on the bacterial surface and provides resistance to phagocytosis [70]. Experiments involving mutants of CPS I synthetic pathway have shown it is required for survival and growth in serum with mutants deficient in CPS I biosynthesis being severely attenuated in hamster models [71]. CPS III is also required for virulence with mutants being attenuated in mouse models [72] although its function is not known. CPS II is upregulated in nutrient deprived conditions suggesting a role in survival in the environment but it is not required for virulence in hamster models [69]. It is unclear what if any role CPS IV plays in pathogenicity or environmental survival.

2.4 Lipopolysaccharide

Lipopolysaccharide (LPS) is the major component in the outer cell membrane of gram negative bacteria. It consists of a lipid A component covalently linked to polysaccharide moiety which varies between bacterial species. *Burkholderia pseudomallei* possesses a unique LPS structure [73] with an unbranched repeating unit of alternating glucose and talose

residues in the form -3)-β-D-glucopyranose-(1–3)-6-deoxy-α-L-talopyranose-(1- with the talose subunits acetylated on the 2' and 3' positions [67]. Mutants in the LPS synthetic pathway are attenuated in mouse, hamster, guinea pig and rat models with the bacteria becoming particularly susceptible to the alternative complement pathway, and cationic microbial peptides [74]. Intact LPS is also required for bacteria survival in nutrient deprived aqueous conditions [75].

2.5 Biofilm formation

Burkholderia pseudomallei are able to form biofilms [76] although different strains have different abilities to form them and production does not correlate with virulence [77]. Cells grown under biofilm inducing conditions have an increased antibiotic resistance although this appears to be due to other genes regulated under the same conditions [78]. Mutants unable to produce usual biofilms are not attenuated in mouse models showing it is not required for virulence [77]. However the presence of bacteria in biofilms in infected lung tissue suggests a role in persistence, not only in harsh environmental conditions, but also in allowing survival inside the host. Biofilm production may therefore partially account for the ability of *Burkholderia pseudomallei* to exist as a latent infection for prolonged periods of time.

2.6 Antibiotic resistance

Burkholderia pseudomallei are resistant to a wide array of antibiotics with a number of antibiotic resistance mechanisms. The *Burkholderia pseudomallei* genome encodes for seven β-lactamases including five Ambler class B (BPSL0374, BPSL1561, BPSL2708 BPSS1915 and BPSS2119) and one of each Ambler class A (BPSS0946) and Ambler class D (BPSS1997) β-lactamase genes providing resistance against a range of beta lactam based antibiotics. There is a putative aminoglycoside acetyltransferase (BPSS0262) encoded in the genome possibly responsible for some of the bacteria's resistance to aminoglycosides alongside other systems.

2.6.1 Efflux pumps

There are a number of multi-drug efflux pumps encoded in the genome sequence including ten members of the RND family. There are seven putative systems (BPSL0307 – BPSL0309, BPSL1269 – BPSL1266, BPSL1566 – BPSL1569, BPSL2234 – BPSL2235, BPSL2871 – BPSL2872 and BPSS1041 – BPSS1043) of unknown function and three fully characterised systems, AmrAB-OprA (BPSL1802 – BPSL1805), BpeAB-OprB (BPSL0813 – BPSL0816)

and BpeEF-OprC (BPSS0290 – BPSS0294) [79]. The AmrAB-OprA efflux pump is specific for the extrusion of aminoglycosides and macrolides [80]. The BpeAB-OprB multi-drug efflux pump is also responsible for the extrusion of aminoglycosides and macrolides [52] and a number of clinical isolates found to be sensitive to aminoglycosides contained mutants in this system [81]. However a separate study on the BpeAB-OprB system from a different strain of *Burkholderia pseudomallei* found no extrusion of aminoglycosides but instead macrolides, fluoroquinolones, tetracyclines, acriflavine, and, to a lesser extent, chloramphenicol [82]. The BpeEF-OprC system is involved in the efflux of chloramphenicol and trimethoprim [83]. There is also a homolog of the NorM multidrug efflux pump (BPSL2468) in the genome. While this is uncategorised in *Burkholderia pseudomallei* it has been shown to be responsible for polymixin efflux in the closely related *Burkholderia vietnamiensis* [84].

2.7 Intracellular survival

Burkholderia pseudomallei is able to inhabit both phagosomal and non-phagosomal cells and encodes several genes involved in intracellular survival within its genome. The bacteria are resistant to human defensin proteins found in epithelial cells when exposed *in vitro* [85]. *Burkholderia pseudomallei* is also able to inhibit DNA and protein synthesis in host cells using at least one uncategorised secreted exotoxin [86]. The environment inside phagocytic cells is challenging, particularly inside phagosomes which the bacteria are able to survive inside. The bacteria have mechanisms which enable it to be highly resistant to reactive oxygen and nitrogen intermediates produced inside macrophage and neutrophil cells, such as superoxide anion, hydrogen peroxide, hydroxy radicals, singlet oxygen and nitric oxide. One method to protect against reactive oxygen species is the production of superoxide dismutase enzymes that catalyse the conversion of superoxide into hydrogen peroxide, which can then be further detoxified by the production of catalse enzymes. The K96243 genome sequence contains two superoxide dismutase genes, a Fe^{2+} dependent *sodB* gene (BPSL0880) and Cu^{2+}/Zn^{2+} sodC (BPSL1001). SodB is currently uncategorised while the SodC protein has been confirmed as possessing superoxide dismutase activity and mutants in this gene are more susceptible to extracellular superoxide [87]. Mutants deficient in SodC also showed decreased survival in macrophage cell lines and were attenuated in BALB/c mouse models of infection [87]. The genome also encodes four catalase genes, three of which have been categorised katG (BPSL2865), katB (BPSS0993) and katE (BPSS2214) and a further putative catalase (BPSL0071). *Burkholderia pseudomallei* can produce a non-specific DNA binding

protein, dpsA (BPSL2863) capable of binding to and protecting DNA from oxidative stress [88]. The genome also encodes a putative hmpA gene (BPSL2840), which is important in resisting oxidative stress in other bacterial species. *Burkholderia pseudomallei* are also resistant to killing by reactive nitrogen species. The genome encodes two alkyl hydroperoxide reductase genes (BPSL1264 and BPSS0492) which have a peroxynitritase activity protecting against reactive nitrogen species [89]. Many of the genes involved in response to intracellular stress are regulated by the alternative sigma factor RpoS. OxyR is also involved in the regulation of oxidative stress response and alongside RpoS, is required for the expression of KatG and DspA [59]. HmpA and SodB were found to be significantly differentially expressed between primary and relapse clinical isolates from melioidosis patients with greater levels found in relapse isolates showing the importance of these mechanisms for long term survival in the host [25].

2.8 Flagella

Burkholderia pseudomallei possess a polar tuft of two to four flagella that provide motility; however its role in virulence is unclear with conflicting data. Mutants deficient in fliC, the flagellin gene, have been observed as significantly less invasive into cultured human lung cells [90]. However a separate study using a different initial strain of the bacteria found fliC mutants were able to invade and replicate inside cultured human lung cells normally but were avirulent in mouse models when introduced intranasally but not intraperitoneally [91]. Similarly fliC mutants are not significantly attenuated in hamster or rat models when introduced intraperitoneally [92]. These data suggest that the flagella may play a role in initial epithelial attachment but deficiencies in flagella production can be overcome by other means.

2.9 Actin motility and multi nucleated giant cell formation

Once inside a host cell *Burkholderia pseudomallei* are able to induce continuous actin polymerisation from one pole of the cell providing intracellular motility. The bacteria uses a novel Arp2/3 independent method mediated by BimA (BPSS1492) [31, 93]. BimA locates to the bacterial surface at one cell pole and stimulates polymerisation of actin by acting as a nucleation promoting factor mimic [31]. Actin polymerisation can result in membrane protrusions which allow for cell to cell spread, with the bacteria entering neighbouring cells contained within a secondary phagosome [94].

Burkholderia pseudomallei are able to induce the formation of multinucleated giant cells although the mechanism by which this occurs is poorly understood. The bacteria are able to produce multinucleated giant cells both in phagocytic and non-phagocytic cell lines and have been observed in infected tissues from melioidosis patients [94-97]. Actin polymerisation mobility is important in the formation of multinucleated giant cells [94, 96]. In macrophage cell lines multinucleated giant cell formation has been shown to be similar to osteoclastogenesis [95]. This process is mediated by the production of LfpA (BPSS2074) [95].

2.10 Pili and fimbriae

The genome of *Burkholderia pseudomallei* K96243 contains eleven genes encoding type IV pili proteins (BPSL0782, BPSL1821, BPSL1899, BPSL2752, BPSL2756, BPSL3008, BPSL3170, BPSS1593, BPSS1595, BPSS2185, and BPSS2186) including the pillin gene (BPSL0782). The type IV pili is involved in microcolony formation which leads to increased association with cultured human cell lines [98]. Pillin deletion mutants are less virulent in nematode and mouse models as well as displaying reduced adherence to human epithelial cells *in vitro* [99]. The genome also encodes six type I fimbriae (BPSL1007 – BPSL1008, BPSL1626 – BPSL1629, BPSL1799 – BPSL1801, BPSL2026 – BPSL2031, BPSS0091 – BPSS0094 and BPSS0120 – BPSS0121) although no fimbriae have been observed in electron micrographs of the bacterium [100] and any role in pathogenicity is unknown.

2.11 Secretion systems and effectors

Burkholderia pseudomallei are able to produce and secrete a number of proteins with a role in pathogenicity including lipases, proteases, haemolysins and siderophores [101].

2.11.1 Proteases and lipases

The *Burkholderia pseudomallei* genome encodes a number of secreted proteases although their role in virulence appears to be minor. Different strains of *Burkholderia pseudomallei* exhibit different level of proteolytic activity in cell free supernatant, however there is no correlation between protease activity and virulence in mouse models of infection [102]. MprA (BPSS1993) is a serine metalloprotease [103]; the protein is produced as a pro-enzyme that is autoproteolytically processed to produce the active form of the enzyme [104]. The protease is able to do extensive damage to cell lines however; MprA is not required for virulence in mouse models [105]. A second calcium-dependent serine protease has also been

characterised [106]. This protease causes localised tissue damage and necrosis if injected into guinea pig or rabbit models [106, 107]. *Burkholderia pseudomallei* can also produce at least one collagenase enzyme [108]. The genome also encodes further uncharacterised proteases with homologs in pathogenic bacteria species including a homolog of *Pseudomaonas aeruginosa* LasA elastase (BPSL0624) and MucD Ser protease (BPSL0808). The *Burkholderia pseudomallei* genome encodes two characterised (BPSL0338 and BPSL2403) and one uncharacterised (BPSS0067) phospholipase C enzymes which are secreted via the twin arginine translocase general secretory system. BPSL0338 and BPSL2403 are non-hemolytic with neither being required for virulence; however BPSL2403 was found to be involved in cytotoxicity of HeLa cells [109]. In contrast mutants deficient in the uncategorised BPSS0067 were severely attenuated in hamster models [110].

2.11.2 Siderophores

Burkholderia pseudomallei is able to produce siderophores to increase iron uptake into the cell [111]. One system involves the biosynthesis of malleobactin by mbaA (BPSL1776) and mbaF (BPSL1774) with fmtA (BPSL1775) transporting the siderophore across the membrane [112]. Malleobactin is able to remove iron from both transferrin and lactoferrin acquiring it for the bacteria [113]. Putative hydroxamate (BPSL1779 – BPSL1774) and pyochelin (BPSS0581 – BPSS0588) biosynthetic genes have also been annotated in the genome [114].

2.11.3 Type II secretion system

The general secretory system (T2SS) is responsible for secreting a number of the effector proteins from the cell including several of the proteases, lipases and phospholipase C proteins [115]. However mutants for the secretory system are not seriuosly attenuated in hamster models suggesting these proteins play only a minor role in pathogenicity [102].

2.11.4 Type III secretion system

The *Burkholderia pseudomallei* genome encodes for three separate type III secretion systems (T3SS), T3SS1 (BPSS1390 – BPSS1409) and T3SS2 (BPSS1592 – BPSS1630) are homologous to systems found in *Ralstonia solanacearum*, a plant pathogen, and T3SS3 (BPSS1534 – BPSS1553) which is homologous to one found in pathogenic *Salmonella* species. T3SS1 and T3SS2 are involved in the infection of tomato plants with deletion mutants being seriously attenuated [14]. An initial study using deletion mutants of the SctU subunit, a major component of the inner membrane assembly, for the three systems found

only T3SS3 to be an important pathogenicity determinant in mouse and hamster models [116]. T3SS3 has been shown to have a role in survival and persistence inside macrophages, escape from endocytic vesicles, multinucleated giant cell formation and induction of host cell apoptosis [117] [118]. More recently a role for T3SS1 has been suggested in mice models using SctN (BPSS1394) deletion mutants with the system being involved in preventing usual phagosomal maturation [119]. Three proteins, BipB (BPSS1532), BipC (BPSS1531), BipD (BPSS1529), form the T3SS3 tip complex required for translocation of proteins across the host cell membrane. The tip complex is required for multinucleated giant cell formation and induction of host cell apoptosis [118] and mutants are attenuated in mice models [120]. Four type III secretion system effector molecules have been identified to date, BopA (BPSS1524), BopC (BPSS1516), BopE (BPSS1525) and Cif (BPSS1385). The BopA protein has been shown to be involved in avoidance of autophagy [121]. BopE acts as a guanidine nucleotide exchange factor for host cell Rho family of GTPases inducing host cell ruffling and actin rearrangements [122]. The T3SS3 effector proteins, BopA and BopE have been shown to exhibit effects in cell lines but deletion mutants do not appear to be significantly less virulent in animal models of infection [116]. This could be due to each molecule playing only a small role but together, alongside other virulence factors, the combined effects of these proteins could be large. BopC represents another T3SS3 effector protein with deletion mutants being less invasive into epithelial cells [123] although experiments to determine the pathogenicity of deletion mutants have not yet been conducted. The Cif protein targets the eukaryotic cell cycle irreversibly blocking the cycle at the G2/M and G1/S transitions [124] which may be important in the production of multi nucleated giant cells.

2.11.5 Type VI secretion system

There are six type VI secretion system (T6SS) gene clusters in the *Burkholderia pseudomallei* K96234 genome [125]. T6SS I mutants are unable to form multinucleated giant cells, escape from phagosomes, grow inside macrophage cell lines and have reduced cytotoxicity of macrophages [126, 127]. T6SS I mutants are severely attenuated in hamster models [127]. T6SS V expression was upregulated following macrophage invasion [125]. T6SS I mutants were attenuated in the cockroach model of infection with mutations in T6SS II – VI showing no loss of virulence [128]. No secreted effector molecules for any of the T6SSs have been identified to date.

2.12 Toxin

BLF1 (BPSL1549) represents a major toxin in *Burkholderia pseudomallei*. The protein acts as a potent cytotoxin against human cell lines, is lethal when injected into mice and insertion mutants are severely attenuated in mouse models [129]. The toxin inhibits the helicase activity of translation factor eIF4A by a glutamine deamidation reaction preventing protein expression [129]. It is unclear how this toxin is exported from the cell or enters into host cells as it lacks a signal sequence or translocation domain. BLF1 was found to be expressed at a significantly higher level in clinical isolates from primary as opposed to relapsing melioidosis patients [25] suggesting a role in initial infection.

3.0 Project aims, target selection and sequence analysis

This section describes the target selection of a number of proteins of unknown function for study in a small-scale structural genomics project. An initial *in silico* analysis was performed on the selected proteins, to ascertain their suitability and merit for this study.

3.1 Project aims

Developments in DNA sequencing led to the production of vast amounts of data in the form of whole genome sequences. Concurrent advances in the technologies required for structure elucidation, from gene cloning, through protein production and purification, to crystallographic techniques, have been combined to provide a high-throughput approach to elucidating the structural and functional characteristics of the proteins encoded by the sequenced genomic DNA. The aims of this project were to identify potential pathogenicity determinants of unknown function from *Burkholderia pseudomallei*, and elucidate their structures with the intention of ascribing a function based on structural analysis or informed functional studies.

3.2 Target selection

The targets in this study represent potential pathogenicity determinants of unknown function from the bacterium *Burkholderia pseudomallei*. Targets were selected based on two characteristics, differential expression by a pathogenic species and a closely related non-pathogenic species, and proteins whose expression is under the control of the RpoS stress response. BLF1 (BPSL1549), a protein found to represent a major toxin of *Burkholderia pseudomallei* [129], has recently been worked on and characterised within the Sheffield crystallography research group and shares both of these characteristics, validating this target selection strategy.

3.2.1 Proteomic comparison

A comparison of protein expression between *Burkholderia pseudomallei* and the closely related, non-pathogenic bacterium *Burkholderia thailandensis* found fourteen genes that were significantly differentially expressed between the two organisms during stationary phase growth [130] (figure 3.1). Six of these genes were homologs of previously known virulence factors from other bacteria. Four of the remaining genes were annotated as hypothetical proteins of unknown function, BPSL1549, BPSL1958, BPSS0683 and BPSS0212.

Burkholderia pseudomallei ***Burkholderia thailandensis***

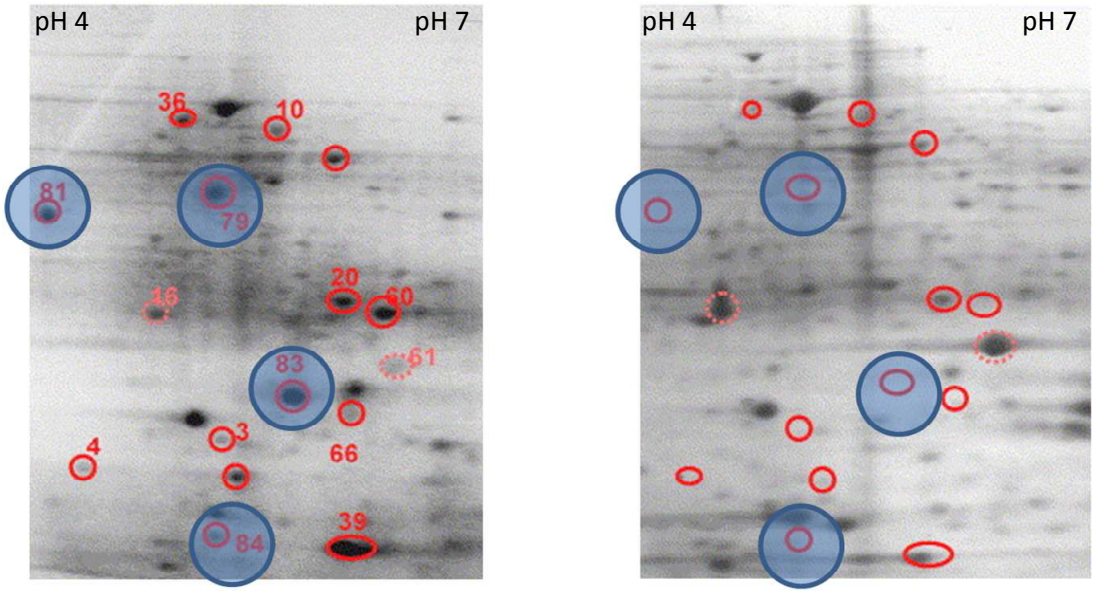


Figure 3.1 2D gel electrophoresis of protein extracts from stationary phase *Burkholderia pseudomallei* and *Burkholderia thailandensis*. The red circled spots represent genes that were highly differentially expressed between the two organisms with the blue circles representing the genes of unknown function that were selected as potential targets for this study. Spot 79 is BPSL1549, spot 81 is BPSL1958, spot 83 is BPSS0212 and spot 84 is BPSS0683. Figure adapted from "Comparative proteomic profiles and the potential markers between *Burkholderia pseudomallei* and *Burkholderia thailandensis*" [130].

3.2.2 Genes under the control of RpoS

A proteomic analysis of $RpoS^+$ and $RpoS^-$ strains followed by a genomic analysis identified 68 genes under direct or in-direct control of RpoS-dependent promoter sites [131]. Eight of these were annotated as hypothetical proteins of unknown function under the direct control of RpoS, BPSL0599, BPSL1549, BPSL3012, BPSS0212, BPSS0213, BPSS0683, BPSS1588 and BPSS2055. BPSS0212 and BPSS0213 exist as part of an operon also containing BPSS0211 and BPSS0214. Along with the four selected genes heavily expressed in *Burkholderia pseudomallei* and not in *Burkholderia thailandensis*, BPSL0599 and BPSS0213 were also identified as being highly expressed during stationary growth but below the threshold for assignment as differentially expressed [132].

3.3 Initial *in silico* analysis of targets

Combined the two approaches to target selection gave eleven unique protein targets, with five sharing both selection criteria. Two of the targets have been studied previously within the Sheffield crystallography group, BPSL1549 and BPSS0683, leaving nine proteins of interest for this study.

3.3.1 Blast and literature search

Blast searches were conducted to find potential homologs in other organisms of known function or structure. BPSL0599 represents a member of the BMA_0021 family of conserved hypothetical proteins found in several *Burkholderia* species. Proteins of this family are often found in thiazole/oxazole-modified microcin biosynthetic clusters, suggesting a possible role in bacteriocin biosynthesis. BPSL1958 homologs are found only in *Burkholderia pseudomallei*, *mallei* and a single strain of *thailandensis* (MSMB43) known to contain genes from other Burkholderia species not normally found in *Burkholderia thailandensis* including the toxin BLF-1. It is unknown if this strain represents a possible human pathogenic strain of *Burkholderia thailandensis*. BPSL1958 contains a highly repetitive sequence of a 52 amino acids repeated six and three quarter times (figure 3.2 a). The repetitive nature is also apparent at the DNA level (figure 3.2 b) suggesting the gene was created in a recent duplication event and there has not been enough time for the sequence of the individual repeats to drift. BPSL3012 is a member of the YacF family of proteins found in many organisms. The function of proteins in this family is unknown but there is a structure in the protein data bank for a homolog from *Vibrio parahaemolyticus* (pdb 2OEZ) [133]. The protein BPSL3012 is known to be associated with the outer membrane or periplasmic space [134] [135]. A BLAST search on the proteins in the BPSS0211 – BPSS0214 operon shows that there are homologs of these proteins in several *Burkholderia* species as well as a number of related plant pathogens. The BLAST search predicts that BPSS0212 and BPSS0213 both contain two separate domains of unknown function, DUF1842 and DUF1843, with the shorter BPSS0211 also representing DUF1843. When the amino acid sequences are aligned it is apparent that BPSS0212 and BPSS0213 are homologs of each other with BPSS0211 representing a homolog of the C-terminal domain of the two proteins (figure 3.3). BPSS0212 and BPSS0213 are associated with the outer membrane or periplasmic space [134] [135] and are also up regulated by RpoE, another alternative sigma factor involved in stress response [57]. The two proteins were found to be significantly differentially expressed between clinical isolates from primary and relapse melioidosis patients suggesting a role in initial infection [25]. BPSS0214 is also of unknown function and is usually, but not always, found associated with the other members of the operon. The protein is predicted to contain a radical-SAM domain, found in a diverse range of proteins involved in processes such as methylation, isomerisation, sulphur insertion, and protein radical formation [136]. BPSS1588 has homologs, all of unknown function, in *Burkholderia mallei*, *thailandensis*, *oaklahomis* and *gladioli* and is also found in *Shewella denitrificans*. BPSS2055 has homologs in *Burkholderia*

Protein	mw (kDa)	pI	Sequence
BPSL0599	13.8	4.9	MAENNAVPTH QSLLDFPEVY LRAIALSWEN EQFKRELLAD PLDALERYFD YRCPWILNLK VAEVPPDESH YGWDAKRQRW NLPVNTLRVG IPTPPKHLAE EGIALAAYND AGPAYLFTCC
BPSL1958	36.1	4.4	MAKLAASNQF GIPNQTDVFA VDGNGSLRVS WVVSAGAWNG PAQIGPAGLF PSRAAVASSN QFGIPNQTDV FAVGRDGALN VAWVVSADRW NGPTPISAAG LFPAGAAIAA SNQFGIPNQT DVFAVSDSGA LNVAWVVSAE RWNGPIPISA AGHFPAGAPL ATSNQFGIPN QTDVFVVDNK GALNVAWVVG AGSWNGPIPI SPPGLFPPGA AVAASNQFGI PNQTDVFVVD NQGALNVAWV VGADRWNGPV PISPAGLFPP GAAVAASNQF GIPNQTDVFA VGRDGALRVA WVVSAGNWNG PVSISPTNLF PSGAAVAASN QFGIPNQTDV FAADSDGVLH VAWVVSAGNW NGPISIA
BPSL3012	28.9	6.2	MILYEYPFNE RIRTLLRLED LFERFTFFVA QEDAREHHVA LTTLFEISEV AGRADLKSDL MKELERQRQT LAPFRGNPGI EQNALEAVLG EIEQTLANLA QMQGKTGQHL IDNEWLASIR SRAVIPGGTC KFDLPSYYAW QQWPAEQRRH DIAKWAMPLL PLRDAAMIVL RLARESGQAS KVMAMQGSYQ QMLSGRTYQL MQVRVPPELR VIPEASANKY MLWVRFTAQD GDVRPRAVDI DVPFQLTLCN L
BPSS0211	6.8	7.9	MSQAQGHPVT PYGVAIHQAI ADGDLAQMKS LRTQAQALLA QQGNLATALE LLEVEIAKLE RRK
BPSS0212	22.5	5.3	MSEDLRVGLF PVRYLVGTGL PGAPQLVLDL MVDTVDHSVV GRAAVSQAVS PPLNFHADVW GSYVFRLGPP PRRDGSGAIV QISLQGNQGG PQSNSMITFY GELLLKGDGK TGVASYRYYS NGSWHEVENV PVKADPELVP IEPGPVIGQS SMSAIGSAAM YGVAIQSAAA SGDLAHMRTL SAYARQQLES RDEIAAALSE LKAEIAKLES RQ
BPSS0213	21.7	6.8	MATTGLFPVQ LRVATPNLGA PVLWLNLLVN TVEKTASGFA RITQTVYPPM HFRARVVGPF HQMRIDPHAP QSVTLTLSGS PTGPVAPQVV ILELNALLNE GWQSGTANYR YFYESRWHSI EHAIVSKDNS RIPLDPPSEH VMPMYGVGLQ EARASGDLSR MKALAQQAEQ QLADHDVIAA ELQKLEAEIA RLEARR
BPSS0214	53.7	7.5	MSTTDARPAR YLFDSDYQRF VPVHAVWEIT LACDLKCLHC GSRAGHRRTN ELSTAECLEV IDALARLGTR EVSLIGGEAY LRKDWTQLIR AIRSHGMYCA IQTGGRNLTP KRLAQAVDAG LNGVGVSLDG LAPLHDKVRN VPGAFERALD TLRRARDAGI AVSVNTQIGA QTMEDLPALM DTIIELGATH WQIQLTVAMG NAVDNDELLL QPYRLAELMP LLAKLYKDGV SRGLLMTVGN NIGYYGPYEH LWRGFGDERV HWSGCAAGQN VIALEADGTV KGCPSLATVG FSGGNVRDMS LEDIWRTSEG IHFGRLRSVD DLWGFCRTCY YADVCRGGCT WTSHSLLGKP GNNPYCHYRV LELQKQGLRE RIAKVQDAGP ASFAVGRFDL VTERIADGEP VASVVRSGQV IELAWKNRGK RSPEVGRVPP KLKMCRNCDG YVHAGEQTCP HCGGDIDAAA RAHELDAQRR HALMNDLERL LGLPASTFGG G
BPSS1588	49.3	6.2	MKKLKYAAAL LTAVAMSPSW SQASTLVAQS HVDGVSRAEP AKIGEQLAAR RASLPTLPRP LPTLSAGAIR QAGLRAPLAK RQATLAAPAA ATVAPPVANC TDVTIGAAYN AATAPAGQAD CFQFVAPSAT KIVAYVVNLP ANEQHDAHLV QVNEDGSWTV LDSQADLSPN KIVEAVPNGP VRLLLLVSAQ QGAGNAPFQF QVLGTTGYDS YEPNDSILHP TKLTGNQLIS ANLDTVADFD YYAVQVPSTQ TANYVTFKGA GTQTAELETA PNTWATLASG TSYNITSPAG ATLMFRVYDK GTTAPAAQAY TLRISDGAGT AGFYRFLDEE NITHLVRGNE NVARVVSAGT IAWDSTGNVR LPPGERIWLR AYDSAGPNGP NTLLSETSGY TDANGNLLVN LNVGVCQGGG TMTGDFNTMS VPSDRWRITY NPYAFVVAYL DNAQIRAQTS IKHFTHICTE QYLGRK

BPSS2055	47.2	6.1	MQNWQPTNPV ERRFMDAHAD WMRFAKDPAA RLMIWQTDET DAQLVQLYFQ GQEETSCAVR TMRARFVNEA RYAQALTDEL VAFYDSRREA SYAQGLQADW QAPPRNDGAS PVLHLLTVAD SLMRHHPDIF PAMVFVLEPA KVRDDAAWVQ WLDGLLSIVA ASPSLGERVR FVVPRTDATP LAALLQRHPD SVRVVHGRYS MASVPRELLA ESGERGPSGE FRRLFVMLTE TIEGGSPARL EELRAAALKV AEREQWFDQC VVVHLIAGAA YLKWRDRERA IEAYRSAADS GMRAVEAGHP AGHKLVANGL FGEASVHLTH KDFARSAYCY ERAAASTSAA QDALMTVEAW RMSAVCWEKA GEREHALEAG FNALDAGLTI DESMRMNSNL RPVVEWMVSQ VGAFDRRRDK LSEKVAALRG GR

Table 3.1 Physical properties and sequences of target proteins.

Protein	Start	End	Length	PFAM Ascension	Name	bit score	E-value
BPSL3012	16	225	211	PF07072.6	DUF1342	267	9.60E-80
BPSS0211	10	62	53	PF08898.5	DUF1843	89.1	1.20E-25
BPSS0212	7	124	114	PF08896.5	DUF1842	139.7	2.60E-41
BPSS0212	159	211	53	PF08898.5	DUF1843	86	1.10E-24
BPSS0213	4	117	114	PF08896.5	DUF1842	133.6	2.20E-39
BPSS0213	143	195	53	PF08898.5	DUF1843	86.6	7.40E-25
BPSS0214	27	183	166	PF04055.16	Radical SAM	76.7	2.00E-21

Table 3.2 Predicted domain architecture of target proteins.

mallei, *thailandensis*, *oaklahomis* and *ubonensis* all of unknown function. It is also found in a variety of other bacterial species including the pathogenic *Yersinia pseudotuberculosis* and *Yersinia pestis*.

(a) Amino acid sequence alignment

```
Repeat1      1 MAKLAASNQFGIPNQTDVFAVDGNGSLRVSWVVSAGAWNGPAQIGPAGLFPS
Repeat2      1 RAAVASSNQFGIPNQTDVFAVGRDGALNVAWVVSADRWNGPTPISAAGLFPA
Repeat3      1 GAAIAASNQFGIPNQTDVFAVSDSGALNVAWVVSAERWNGPIPISAAGHFPA
Repeat4      1 GAPLATSNQFGIPNQTDVFVVDNKGALNVAWVVGAGSWNGPIPISPPGLFPP
Repeat5      1 GAAVAASNQFGIPNQTDVFVVDNQGALNVAWVVGADRWNGPVPISPAGLFPP
Repeat6      1 GAAVAASNQFGIPNQTDVFAVGRDGALRVAWVVSAGNWNGPVSISPTNLFPS
Repeat7      1 GAAVAASNQFGIPNQTDVFAADSDGVLHVAWVVSAGNWNGPISIA-------
consensus    1 .*..*.*************...  *.*.*.***.*. ****..*.......
```

(b) DNA sequence alignment

```
Repeat1      1 ATGGCAAAATTAGCAGCTTCAAATCAATTCGGCATTCCCAATCAGACCGACGTATTTGCC
Repeat2      1 CGCGCGGCGGTTGCGTCGTCGAACCAATTCGGCATCCCGAATCAGACCGACGTGTTCGCC
Repeat3      1 GGCGCGGCGATCGCGGCGTCGAACCAGTTCGGCATCCCGAATCAGACCGACGTGTTTGCC
Repeat4      1 GGCGCGCCGCTCGCGACATCGAACCAGTTCGGCATCCCGAATCAGACTGACGTGTTCGTC
Repeat5      1 GGCGCGGCGGTCGCCGCATCGAACCAGTTCGGCATTCCGAATCAGACCGACGTGTTCGTC
Repeat6      1 GGCGCGGCGGTTGCGGCATCGAACCAGTTCGGTATCCCGAACCAGACCGACGTGTTTGCC
Repeat7      1 GGCGCGGCGGTTGCGGCATCGAACCAGTTCGGCATCCCGAATCAGACCGACGTGTTCGCC
consensus    1 ...**.....* **..*.**.**.**.*****.**.**.**.*****.*****.**.*.*

Repeat1     61 GTTGACGGCAACGGCTCGCTGCGCGTTTCCTGGGTGGTCAGCGCCGGCGCCTGGAACGGC
Repeat2     61 GTCGGCCGCGACGGCGCGCTGAACGTCGCGTGGGTGGTCAGCGCCGATCGCTGGAACGGA
Repeat3     61 GTGAGCGACAGCGGCGCGTTGAACGTGGCGTGGGTCGTCAGCGCCGAGCGCTGGAACGGA
Repeat4     61 GTCGACAATAAAGGCGCGTTGAACGTGGCATGGGTTGTCGGCGCCGGCAGTTGGAACGGC
Repeat5     61 GTCGACAACCAGGGCGCGTTGAACGTCGCCTGGGTCGTCGGCGCGGATCGCTGGAACGGG
Repeat6     61 GTCGGCCGCGACGGCGCACTACGCGTTGCCTGGGTGGTCAGCGCCGGCAATTGGAACGGC
Repeat7     61 GCCGATAGCGACGGCGTCCTGCACGTCGCGTGGGTGGTCAGCGCCGGCAATTGGAACGGG
consensus   61 *..........***....*...***  .*.*****.***.****.*.....********.

Repeat1    121 CCCGCTCAGATCGGGCCGGCGGGGCTCTTCCCGTCC
Repeat2    121 CCGACCCCGATCAGCGCCGCGGGCCTTTTCCCGGCC
Repeat3    121 CCGATTCCAATCAGCGCGGCGGGCCACTTCCCGGCG
Repeat4    121 CCGATTCCGATCAGCCCTCCCGGGCTGTTCCCTCCC
Repeat5    121 CCGGTTCCGATCAGCCCTGCCGGGCTCTTCCCGCCC
Repeat6    121 CCCGTGTCGATCAGCCCGACGAACCTTTTCCCGTCC
Repeat7    121 CCGATTTCCATTGCGTGA----------------
consensus  121 **.......**............ .. ...... ..
```

Figure 3.2 Sequence alignments for BPSL1958. **a** Amino acid sequence alignment of the repeated elements of BPSL1958. Identical residues are shaded black with similar amino acids shaded grey. **b** DNA sequence alignment of the seven repeated elements of BPSL1958. Both alignments have been produced using ClustalW [134] and BOXSHADE [135] and the consensus is shown. Residues that are completely conserved are shown as (*), more than half shown as (.) and less than half left blank.

```
BPSS0211      1 ------------------------------------------------------------
BPSS0212      1 MSEDLRVGLFPVRYLVGTGLPGAPQLVLDLMVDTVDHSVVGRAAVSQAVSPPLNFHADVW
BPSS0213      1 ---MATTGLFPVQLRVATPNLGAPVLWLNLLVNTVEKTASGFARITQTVYPPMHFRARVV
consensus     1         .....   ...   ... . . ... ..... . . ... . ... ... .

BPSS0211      1 ------------------------------------------------------------
BPSS0212     61 GSYVFRLGPPPRRDGSGAIVQISLQGNQGGPQSNSMITFYGELLLKGDGKTGVASYRYYS
BPSS0213     58 GPFHQMRIDPH----APQSVTLTLSGSPTGPVAPQVVILELNALLNEGWQSGTANYRYFY
consensus    61 . .       .          . ... .   ..    ..       ..     .. . ....

BPSS0211      1 ----------------------------MSQAQGHPVTPYGVAIHQAIADGDLAQMKSL
BPSS0212    121 NGSWHEVENVPVKADPELVPIEPGPVIGQSSMSAIGSAAMYGVAIQSAAASGDLAHMRTL
BPSS0213    114 ESRWHSIEHAIVSKDNSRIPLDPP---------SEHVMPMYGVGLQEARASGDLSRMKAL
consensus   121    .. ..   .  .   .....         .    . . .***... * *.***..*..*

BPSS0211     32 RTQAQALLAQQGNLATALELLEVEIAKLERRK
BPSS0212    181 SAYARQQLESRDEIAAALSELKAEIAKLESRQ
BPSS0213    165 AQQAEQQLADHDVIAAELQKLEAEIARLEARR
consensus   181   .* ..*. .. .*..*  *..***.** *.
```

Figure 3.3 Amino acid sequence alignment of BPSS0211, BPSS0212 and BPSS0213. Identical residues are shaded black with similar amino acids shaded grey. The alignment has been produced using ClustalW [134] and BOXSHADE [135] and the consensus is shown. Residues that are completely conserved are shown as (*), more than half shown as (.) and less than half left blank.

3.3.2 Primary sequence prediction

The SignalP program [139] was used to predict if any targets had a potential signal sequence. One target, BPSS1588, was predicted to have an N-terminal signal sequence for export from the cell (figure 3.4). This was taken into account when designing primers to create a construct without the signal sequence.

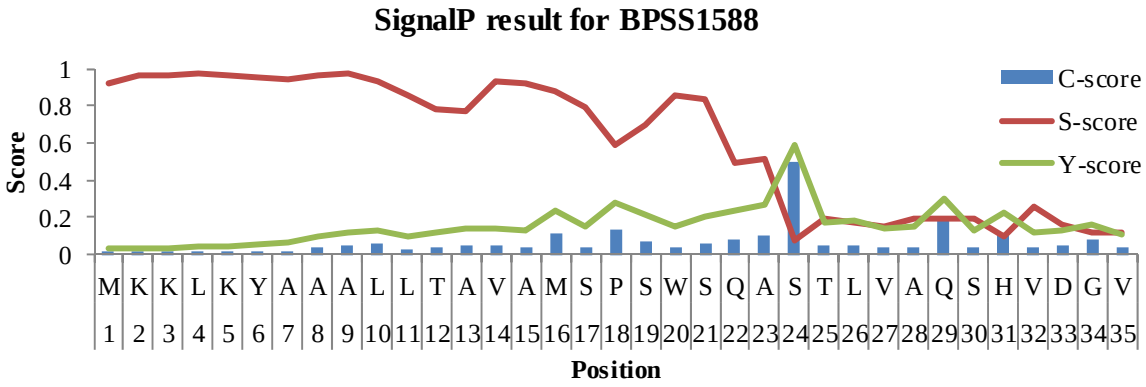
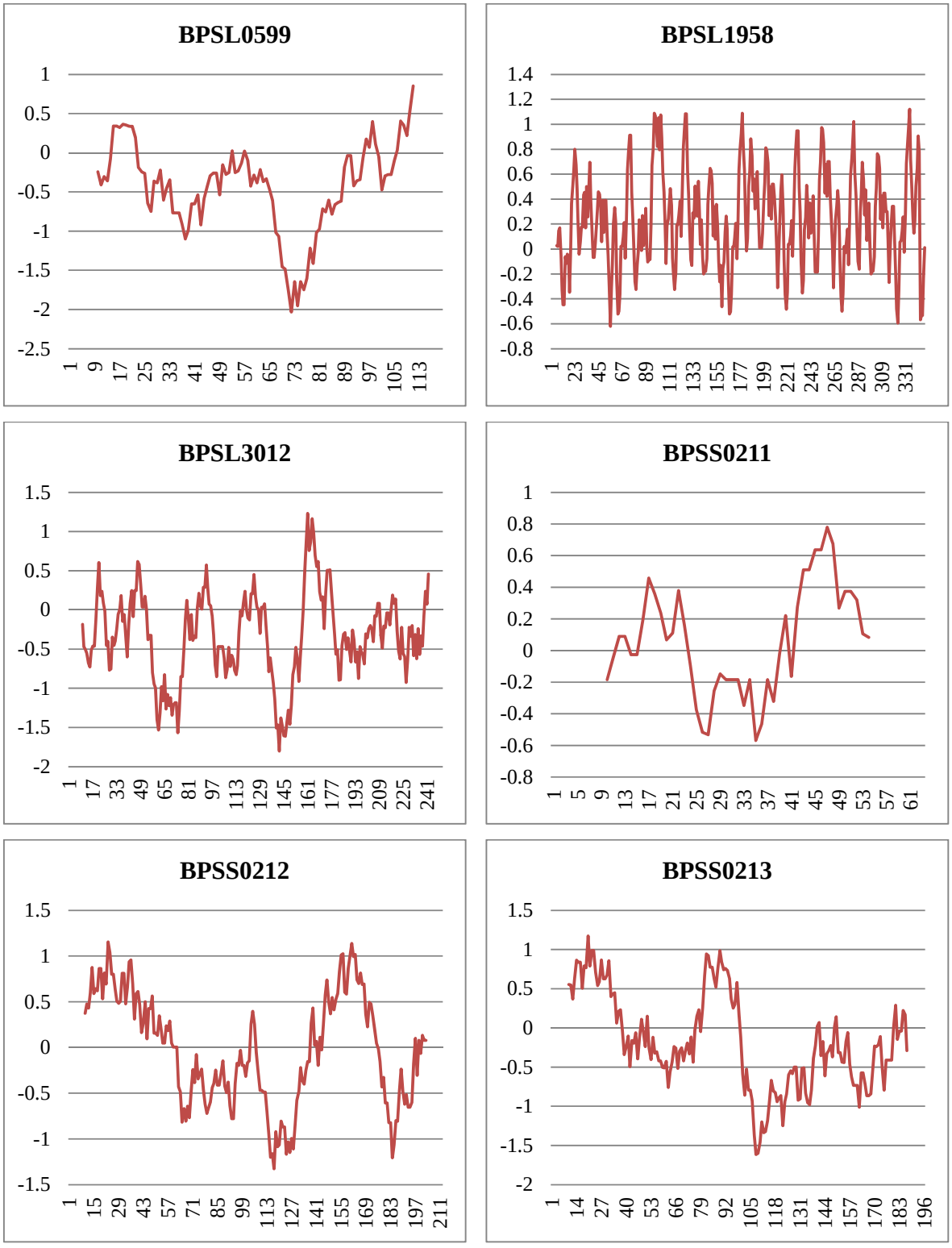


Figure 3.4 Signal sequence prediction for BPSS1588. The prediction was conducted using the SignalP programme [136]. The S-score is based on the likelihood of any residue appearing in a signal sequence while the C-score is the cleavage site score predicting a position where cleavage is likely to occur. The Y-score represents a combination of the two other scores giving the most accurate estimate of the cleavage site in this case after alanine 23 in the protein.

Hydropathy plots calculated for the target genes using Protscale [140] to predict if the proteins were all theoretically soluble suggested none of the targets contained integral transmembrane helicies (figure 3.5).



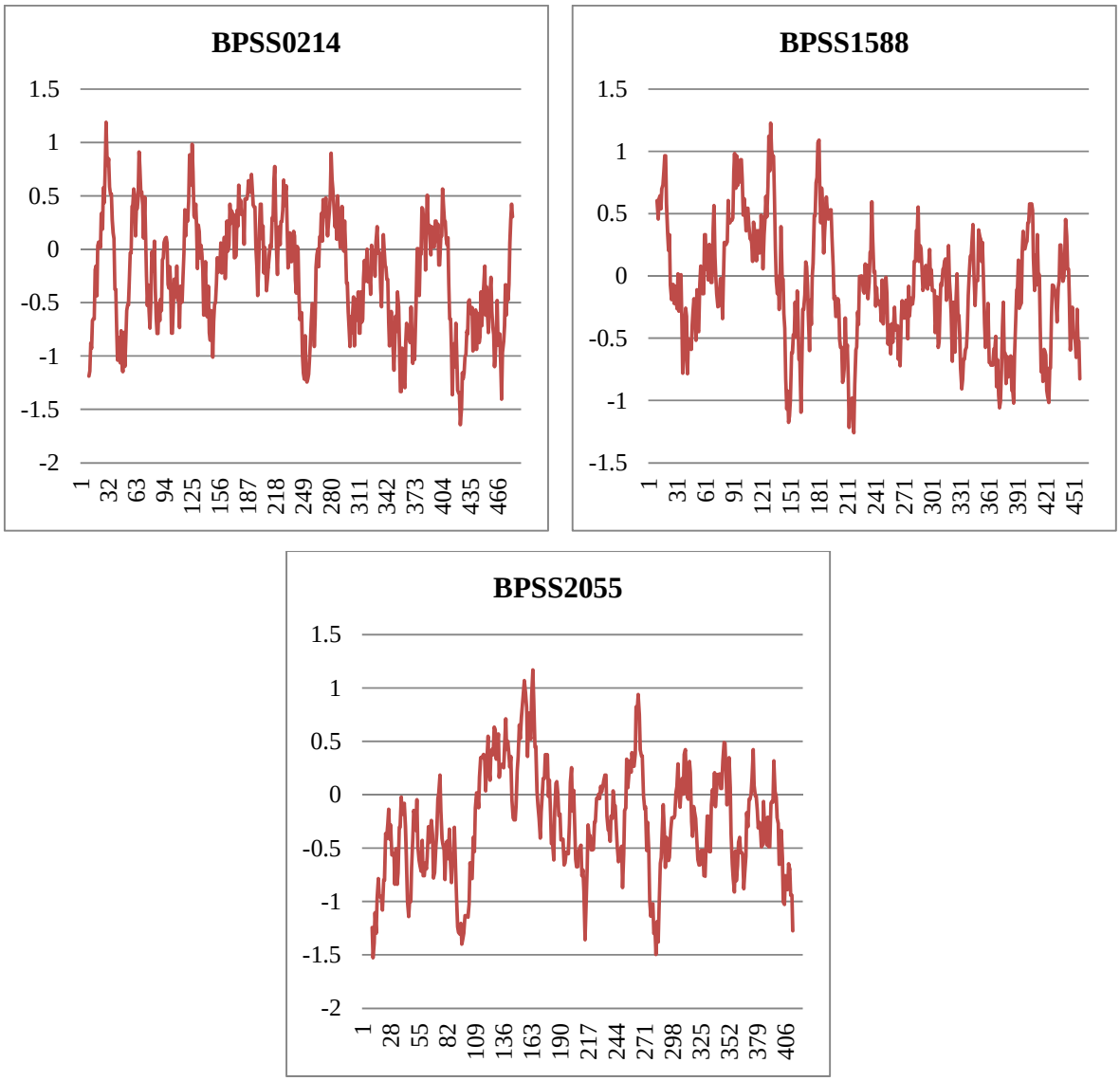


Figure 3.5 Hydropathy plots for the target genes from *Burkholderia pseudomallei*. The figure was created using Protscale [137] with a window size of 19 residues and the Kyte and Doolittle hydrophobicity scores. A value of 1.5 or over would indicate a possible transmembrane helix. As this threshold value is not reached for any of the targets it suggests all target proteins do not contain a membrane spanning helix.

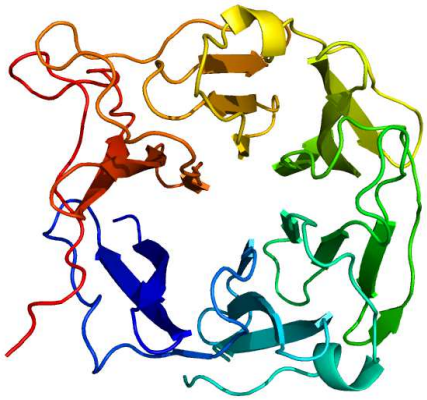
3.3.3 Secondary structure analysis and threading

Secondary structure prediction and threading were also carried out on the target proteins using the Phyre 2 server [141]. The results were largely uninformative, with the exception of BPSL1958, BPSL3012 and BPSS1588. Secondary structure prediction on BPSL1958 showed an abundance of beta sheet in the protein with no predicted helices. Threading analysis predicts the protein folding into a six (figure 3.6 a) or seven (figure 3.6 b) bladed beta-propeller with each of the repeats forming a single blade. BPSL3012 was predicted to have a

similar secondary structure and fold to the homologous protein, YacF, in the PDB (figure 3.6 c). The threading results for BPSS1588 were inconclusive for the majority of the protein however they highlighted the presence of a putative CUB-like collagen binding domain within the protein (figure 3.6 d).

(a) BPSL1958 (6 bladed) model

Residues 13 – 316 (357), PDB 1ofz [139]

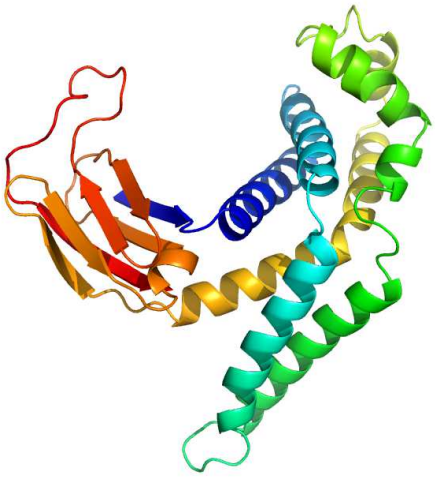


(b) BPSL1958 (7 bladed) model

Residues 9 – 349 (357), PDB 2c4d [140]

(c) BPSL3012 model

Residues 2 – 249 (251), PDB 2oez [131]



(d) BPSS1588 model

Residues 212 – 315 (466), PDB 1nqj [141]

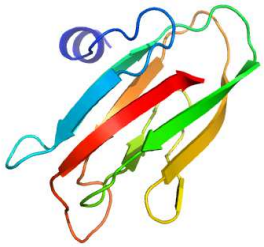


Figure 3.6 PHYRE 2 threading results for BPSL1958, BPSL3012 and BPSS1588. The images show the predicted folds of the proteins. The represented residues of the target protein, with the total number of residues in parenthesis, are listed alongside the PDB entry onto which they have been threaded.

3.4 Structure solution strategy

The overall strategy for structure solution consisted of cloning genes from genomic DNA into a suitable vector for protein expression in *Escherichia coli* followed by purification, crystallisation and data collection. The primary sequence of the target proteins was

considered in order to decide on a possible phasing strategy for structure solution (table 3.3). BPSL3012 has a homolog with 32 % identity in the PDB which would likely allow phasing by molecular replacement. BPSL0599 and BPSS1588 contain a high proportion of tyrosine residues which could allow experimental phases to be obtained using iodine soaking experiments and iodine SAD data collection. BPSS0212, BPSS0213 and BPSS0214 contain a high proportion of methionine residues which would allow phases to be obtained through the incorporation of seleno-L-methionine into the protein and selenium MAD data collection. Alternatively the large number of cysteine residues in BPSS0214 could allow experimental phases to be obtained using mercurial compound soaking and co-crystallisation experiments and mercury MAD data collection. BPSL1958 contains no naturally occurring methionines, cysteines or tyrosines presenting a possible problem for obtaining experimental phase information. The strategy was therefore to create a number of mutants containing either cysteine or methionine residues to enable phasing. For the three homologous proteins BPSS0211, BPSS0212 and BPSS0213 it was hoped that once BPSS0212 or BPSS0213 was solved the others could be phased using molecular replacement.

Protein	**Number of residues**				**Phasing methods**
	Total	**Methionine**[a]	**Cysteine**	**Tyrosine**	
BPSL0599	120	0	3	6	C, **Y**
BPSL1958	357	0	0	0	MR (M, **C**)[b]
BPSL3012	251	9	2	7	**MR**, M, C, Y
BPSS0211	63	1	0	1	**(MR)**[d], **M**, Y
BPSS0212	212	5	0	8	**(MR)**[d], **M**, Y
BPSS0213	196	5	0	5	**(MR)**[d], **M**, Y
BPSS0214	491	9	16	14	MR, **M**, **C**, Y
BPSS1588[c]	443 (466)	3 (4)	4 (4)	17 (18)	M, C, **Y**
BPSS2055	422	14	4	9	**M**, C, Y

[a] This is excluding the N-terminal residue which is frequently removed by the cellular machinery.

[b] These methods are only available following mutation.

[c] Numbers are for truncated protein, those in parenthesis represent the full length protein.

[d] These methods are only available following resolution of a related structure.

Table 3.3 Summary of possible phasing techniques for the target proteins. Possible techniques considered included (MR) phasing using a molecular replacement model, (M) phasing using selo-L-methionine incorporation experiments, (C) phasing using mercurial compound soaking or co-crystallisation experiments, (Y) phasing using iodine soaking experiments. The preferred method(s) for each target is in bold.

Chapter two

Results from a small-scale structural genomics project conducted for proteins from the bacterial pathogen, *Burkholderia pseudomallei*

Section 4 DNA to soluble protein: Producing material for protein studies

Section 5 Studies on the protein BPSL0599

Section 6 Studies on the protein BPSL1958

Section 7 Studies on the protein BPSS0211

Section 8 Studies on the protein BPSS0212

Section 9 Studies on the protein BPSS0213

4.0 DNA to soluble protein: Producing material for protein studies

This section describes the cloning of target genes from genomic DNA into plasmid vectors and the overexpression of target proteins in *Escherichia coli* cells.

4.1 Cloning

Target genes were amplified from *Burkholderia pseudomallei* strain D286 genomic DNA, taken from a melioidosis patient at Kuala Lumpur Hospital, by PCR using BioMix Red (Bioline), varying concentrations of DMSO and specific primers for each gene (Eurofins) (table 4.1). DMSO was included in the PCR reactions in an attempt to improve the production of the desired genes with reduced background contaminants, due to the high GC content of the *Burkholderia pseudomallei* genome [49, 145]. Altering the concentration of DMSO was shown to have an effect on amplification and the conditions resulting in the best amplification without other contaminating fragments were selected (figure 4.1). PCR products were purified using agarose gel electrophoresis and a QIAquick® Gel Extraction Kit (Qiagen) by standard protocols. A miniprep was performed using a QIAprep® Miniprep kit (Qiagen) from DH5α cells containing pET21a plasmid to obtain vector DNA. The purified PCR products and plasmid were then digested using NdeI and either BamHI or EcoRI restriction enzymes (NEB) depending on the site present in the primer sequence (table 4.1). Vector DNA was purified by gel extraction and then treated with antartic phosphatase (NEB), to prevent singularly digested plasmids self-ligating, and inserts were purified using a QIAquick® PCR purification kit (Qiagen). Vector and insert DNA were ligated together using T4 DNA ligase (NEB) before being transformed into competent DH5alpha cells (VWR) and plated on LB-agar with the addition of 100 µg ml^{-1} carbenicillin for selection. Plasmids were recovered by miniprep, sequenced (Geneservice) and compared to the expected sequence from the genome (table 4.2). As different strains of bacteria were used for sequence analysis (K96243) and PCR amplification (D286) a number of mutations were accepted.

4.1.1 Site-directed mutagenesis

BPSL1958 mutants were required in order to enable experimental phases for the structure resolution to be obtained. Sites chosen for cysteine mutagenesis were polar residues (with the exception of the C-terminal alanine), and predicted to be solvent exposed in the tertiary structure of the threading models (figure 4.2), while those for methionine mutagenesis were all isoleucine to minimise the change in amino acid side chain properties. A QuikChange site

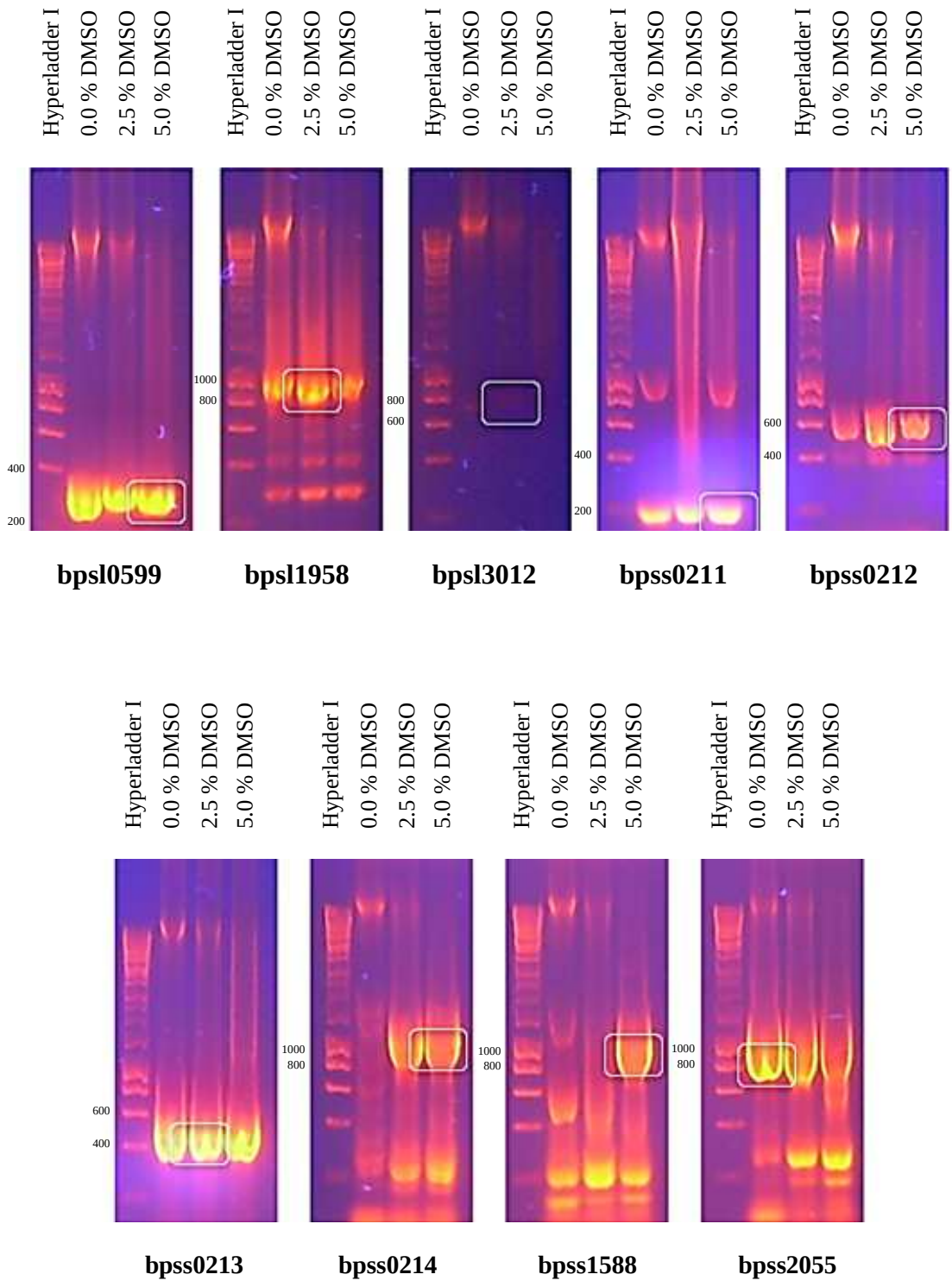


Figure 4.1 Agarose gels showing the amplification of genes by PCR from *Burkholderia pseudomallei* genomic DNA. Circled bands were excised from the gel for purification. bpss2055 amplified best without the addition of DMSO, bpsl1958, bpsl3012 and bpss0213 with the addition of 2.5 % DMSO and bpsl0599, bpss0211, bpss0212, bpss0214 and bpss1588 with the addition of 5 % DMSO.

Oligoname (restriction site)	Sequence
bpsl0599 F (NdeI)	5’-ATATATATCATATGGCCGAGAACAACGCC-3’
bpsl0599 R (EcoRI)	5’-ATATATATGAATTCTCAGCAGCAGGTGAA-3’
bpsl1958 F (NdeI)	5’-ATATATCATATGGCAAAATTAGCAGCTTC-3’
bpsl1958 R (EcoRI)	5’-ATATATGAATTCTCACGCAATGGAAATCG-3’
bpsl3012 F (NdeI)	5’-ATATATATCATATGATTCTTTACGAGTAT-3’
bpsl3012 R (BamHI)	5’-ATATATATGGATCCTTACAGATTGCAGAG-3’
bpss0211 F (NdeI)	5’-ATATATATCATATGAGTCAAGCACAGGGC-3’
bpss0211 R (BamHI)	5’-ATATATATGGATCCCTACTTCCTCCGCTC-3’
bpss0212 F (NdeI)	5’-ATATATCATATGATGTCCGAAGATCTTCG-3’
bpss0212 R (BamHI)	5’-ATATATGGATCCTCACTGCCGGCTTTCGA-3’
bpss0213 F (NdeI)	5’-ATATATATCATATGGCAACTACCGGTCTC-3’
bpss0213 R (BamHI)	5’-ATATATATGGATCCTCAGCGGCGCGCTTC-3’
bpss0214 F (NdeI)	5’-ATATATATCATATGAGCACAACGGACGC-3’
bpss0214 R (BamHI)	5’-ATATATATGGATCCTCACCCGCCGCCGAA-3’
bpss1588 F (NdeI)	5’-ATATATATCATATGTCGACCCTCGTCGCG-3’
bpss1588 R (BamHI)	5’-ATATATATGGATCCTTACTTGCGGCCGAG-3’
bpss2055 F (NdeI)	5’-ATATATATCATATGCAGAACTGGCAGCC-3’
bpss2055 R (BamHI)	5’-ATATATATGGATCCTCAGCGGCCGCCACG-3’

Table 4.1 Primers used for the PCR amplification of genes from genomic DNA designed using the K96243 genome sequence.

Target	Mutations compared to strain K96243 genomic DNA
bpsl0599	no mutations
bpsl1958	S312A[a]
bpsl3012	no mutations
bpss0211	no mutations
bpss0212	Q187R[b]
bpss0213	no mutations
bpss0214	L81Q, E174D and F388S[b]
bpss1588	no mutations
bpss2055	D322G[a]

[a] Multiple plasmids from different colonies produced following transformation were sequenced resulting in the same sequence.

[b] A single plasmid sequence was obtained.

Table 4.2 Summary of sequencing results for target genes. The mutations found could represent PCR errors or a difference in the genomic DNA of the strains used for selection (K96243) and as a template in the PCR reaction (D286). The chance of the mutation arising from PCR error is reduced for genes where multiple constructs were sequenced.

Figure 4.2 Selection of residues for site directed mutagenesis in BPSL1958. The sequence has been split into the seven repeats representing the blades of the predicted β-propeller fold and the secondary structure prediction from the PHYRE2 server is assigned. These individual blades contain four β-strands in the threaded model which are roughly highlighted in orange for each blade. Residues selected for mutation to cysteine residues are highlighted in red and are generally polar amino acids appearing in linking regions between strands. Residues selected for mutation to methionine residues are highlighted in green and are all isoleucine in the original sequence.

directed mutagenesis kit (Stratagene) was used with specific primers (Eurofins) (table 4.3) to create cysteine and methionine mutants. For the first batch of cysteine mutants the wild type gene was used as a template for the mutagenesis. In subsequent reactions to create the methionine and the remaining cysteine mutants, the K3C mutant was used as it allowed for an improved purification strategy (section 6.1.2). The highly repetitive sequence of BPSL1958 created problems during mutagenesis despite sites being carefully chosen based on the relative level of difference between the seven repeats. Colonies were screened by colony PCR using T7F and T7R primers to check the size of the resultant gene before sequencing (Geneservice) to confirm which of the desired mutations had been produced. The following mutants were successfully produced: K3C, S128C, A357C, K3C-D244C, K3C-H340C, K3C-I44M, K3C-I356M, K3C-I44M-I252M and K3C-I44M-I252M-I356M.

Oligoname	Sequence
bpsl1958 K3C F	5’-GGAGATATACATATGGCATGTTTAGCAGCTTCAAATCAATTCGGC-3’
bpsl1958 K3C R	5’-GCCGAATTGATTTGAAGCTGCTAAACATGCCATATGTATATCTCC-3’
bpsl1958 R53C F	5’-CTCTTCCCGTCCTGCGCGGCGGTTGCGTC-3’
bpsl1958 R53C R	5’-GACGCAACCGCCGCGCAGGACGGGAAGAG-3’
bpsl1958 R89C F	5’-GTCAGCGCCGATTGCTGGAACGGACCGAC-3’
bpsl1958 R89C R	5’-GTCGGTCCGTTCCAGCAATCGGCGCTGAC-3’
bpsl1958 S128C F	5’-CCGTGAGCGACTGCGGCGCGTTGAACG-3’
bpsl1958 S128C R	5’-CGTTCAACGCGCCGCAGTCGCTCACGG-3’
bpsl1958 H153C F	5’-GCGCGGCGGGCTGCTTCCCGGCGGGCG-3’
bpsl1958 H153C R	5’-CGCCCGCCGGGAAGCAGCCCGCCGCGC-3’
bpsl1958 K180C F	5’-GACGTGTTCGTCGTCGACAATTGTGGCGCGTTGAACGTGGC-3’
bpsl1958 K180C R	5’-GCCACGTTCAACGCGCCACAATTGTCGACGACGAACACGTC-3’
bpsl1958 D244C F	5’-CGTCGGCGCGTGTCGCTGGAACGG-3’
bpsl1958 D244C R	5’-CCGTTCCAGCGACACGCGCCGACG-3’
bpsl1958 S312C F	5’-CCTTTTCCCGTGCGGCGCGGCGGTTGC-3’
bpsl1958 S312C R	5’-GCAACCGCCGCGCCGCACGGGAAAAGG-3’
bpsl1958 H340C F	5’-GACGGCGTCCTGTGCGTCGCGTGGGTGGTC-3’
bpsl1958 H340C R	5’-GACCACCCACGCGACGCACAGGACGCCGTC-3’
bpsl1958 A357C F	5’-GGCCGATTTCCATTTGTTGAGATCCGAATTCGAGCTCCG-3’
bpsl1958 A357C R	5’-CGGAGCTCGAATTCGGATCTCAACAAATGGAAATCGGCC-3’
bpsl1958 I44M F	5’-CGGCCCCGCTCAGATGGGGCCGGCGGGGC-3’
bpsl1958 I44M R	5’-GCCCCGCCGGCCCCATCTGAGCGGGGCCG-3’
bpsl1958 I64M F	5’-GAACCAATTCGGCATGCCGAATCAGACCG-3’
bpsl1958 I64M R	5’-CGGTCTGATTCGGCATGCCGAATTGGTTC-3’
bpsl1958 I108M F	5’-GGCCGGCGCGGCGATGGCGGCGTCGAACC-3’
bpsl1958 I108M R	5’-GGTTCGACGCCGCCATCGCCGCGCCGGCC-3’
bpsl1958 I252M F	5’-CGGGCCGGTTCCGATGAGCCCTGCCGGGC-3’
bpsl1958 I252M R	5’-GCCCGGCAGGGCTCATCGGAACCGGCCCG-3’
bpsl1958 I356M F	5’-CGGGCCGATTTCCATGGCGTGAGAATTCG-3’
bpsl1958 I356M R	5’-CGAATTCTCACGCCATGGAAATCGGCCCG-3’

Table 4.3 Primers used for the site-directed mutagenesis of BPSL1958 to create cysteine and methionine mutations.

4.2 Overexpression

Plasmids containing correctly sequenced genes were used to transform BL21 (DE3) or Tuner (DE3) competent cells (Novagen) (table 4.4) for overexpression. A small-scale over expression trial was performed for all genes to find post induction conditions that would produce soluble protein. Successful conditions were identified for the production of BPSL0599, BPSL1958, BPSS0211, BPSS0212 and BPSS0213 (table 4.4).

Target	*Escherichia coli* Strain	Temperature (°C)	[IPTG] (mM)	Time (hours)
BPSL0599	BL21(DE3)	20	1.0	3
BPSL1958	BL21(DE3)	9[a]	1.0	30
BPSL1958 A357C, K3C-D244C and K3C-H340C	BL21(DE3)	8[a]	1.0	30
BPSL1958 K3C and S128C	Tuner(DE3)	8[a]	1.0	30
BPSL1958 K3C-I44M-I252M-I356M (Seleno-Met)	Tuner(DE3)	10	1.0	48
BPSS0211	BL21(DE3)	37	1.0	3
BPSS0211 (Seleno-Met)	BL21(DE3)	37	1.0	18
BPSS0212	BL21(DE3)	12[a]	1.0	24
BPSS0212 (Seleno-Met)	BL21(DE3)	12	1.0	48
BPSS0213	BL21(DE3)	15[a]	1.0	24

[a] cold shock prior to induction

Table 4.4 Post-induction conditions for the soluble overexpression of target proteins.

BPSL0599 was found to overexpress as two separate molecular weights. Purified protein was analysed by mass spectrometry (Simon Thorpe, University of Sheffield) showing that the two products were the full length protein and a smaller fragment clipped by the cellular machinery (figure 5.3). The remaining proteins, BPSL3012, BPSS0214, BPSS1588 and BPSS2055 could only be produced as part of the insoluble cellular fraction. Mutant constructs of BPSL1958 overexpressed under the same conditions as for native BPSL1958. Larger scale growths were conducted to obtain enough cell paste from which to purify the proteins of interest. For BPSL1958, BPSS0212 and BPSS0213 it was necessary to adjust the overexpression protocol once the expression had been scaled up. It was required for the

cultures to be incubated on ice for 20 minutes, immediately prior to the addition of IPTG. Without this extra step the proteins were produced primarily in the insoluble fraction.

4.2.1 Seleno-methionine

Seleno-methionine protein was produced for BPSL1958 K3C-I44M-I252M-I356M, BPSS0211 and BPSS0212 in order to obtain experimental phasing. For the production of seleno-methionine incorporated protein, cells were grown in LB media before being transferred to minimal media supplemented with seleno-methionine prior to induction of protein expression. For BPSS0211 and BPSS0212 the same overexpression conditions for native protein were applied with a longer post induction incubation time (table 4.4). For BPSL1958 K3C-I44M-I252M-I356M it was necessary to conduct a further small-scale expression trial using minimal media supplemented with seleno-methionine for suitable conditions to be found for overexpression (table 4.4).

4.3 Summary of cloning and overexpression

The nine initial targets were successfully cloned into pET21a vectors and transformed into expression strains where five were successfully expressed as soluble proteins. Mutant forms of BPSL1958 were produced by site directed mutagenesis and expressed as part of the cells soluble fraction. Seleno-methionine incorporated cell paste was also produced for BPSL1958 K3C-I44M-I252M-I356M, BPSS0211 and BPSS0212. The following five sections deal with subsequent work conducted for the successfully expressed protein targets.

5.0 Studies on the protein BPSL0599

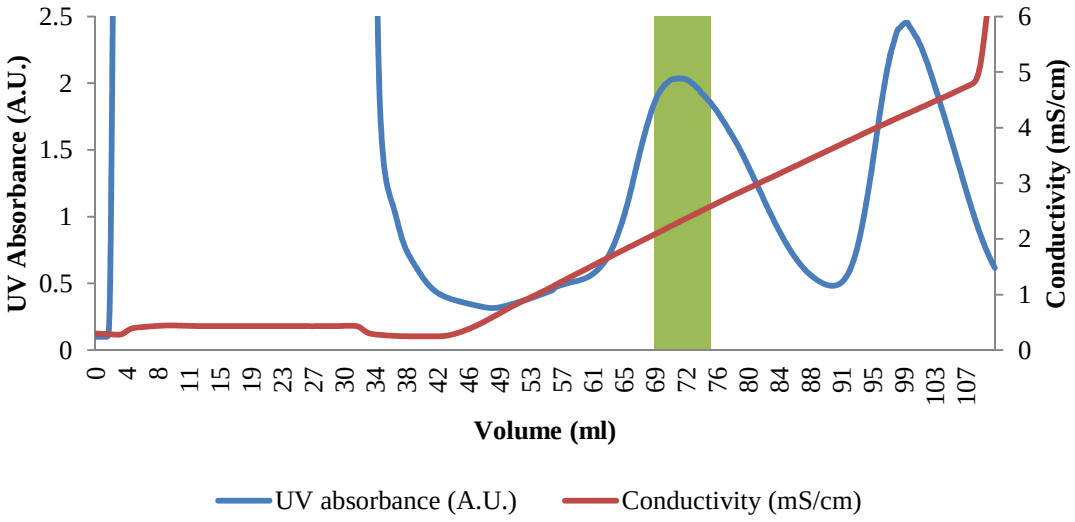
This section describes the purification, and crystallisation of the target BPSL0599.

5.1 Protein purification for BPSL0599

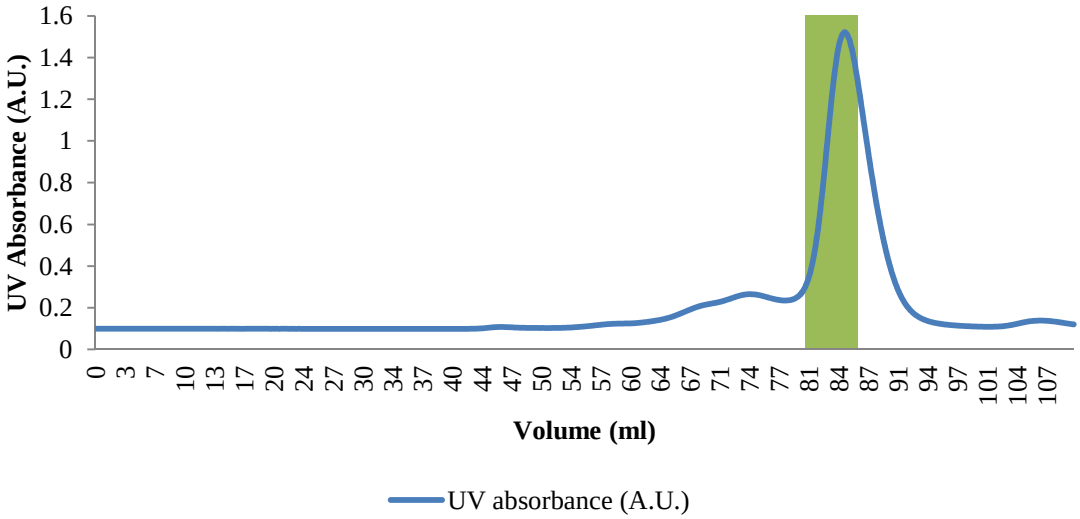
Approximately 2 g of cell paste was resuspended in 30 ml 50 mM TRIS pH 8.0 and disrupted by sonication. Cell debris and insoluble proteins were removed by centrifugation at 70,000 g for 15 minutes and the supernatant was loaded onto a DEAE-HiTrap column equilibrated with 50 mM TRIS pH 8.0. A 75 ml gradient from 0 to 500 mM NaCl was then applied to the column and 2.5 ml fractions were collected (figure 5.1 a). Peak fractions were combined and concentrated to 1.5 ml using a Vivaspin concentrator with a 10 kDa MWCO. This was then loaded onto a 1.6 x 60 cm Superdex 200 gel filtration column equilibrated in 50 mM TRIS pH 8.0 and 500 mM NaCl and eluted using the same buffer collecting 2 ml fractions (figure 5.1 b). Fractions were analysed by SDS-PAGE (figure 5.2) and those containing BPSL0599 were combined. At this stage the sample, which contained a mixture of the two BPSL0599 fragments produced during overexpression (sections 4.2 and 5.1.1), was split with half of the sample being concentrated to 12 mg ml^{-1} and having its buffer exchanged for 10 mM TRIS pH 8.0 for use in crystallisation trials. The remaining sample was diluted 20 fold using 50 mM MES pH 6.3 and applied to a ResourceQ column equilibrated with 50 mM MES pH 6.3. A 120 ml gradient from 0 to 250 mM NaCl was then applied to the column and 2.5 ml fractions were collected (figure 5.1 c) and analysed by SDS-PAGE (figure 5.2). This final step allowed for the two molecular weight components of the sample to be partially separated. The buffer was exchanged for 10 mM TRIS pH 8.0 and the two protein samples were concentrated to 10 mg ml^{-1} for use in crystallisation trials.

Figure 5.1 Chromatogram traces for the purification of BPSL0599. **a** DEAE purification step showing column loading and elution. 2.5 ml fractions were collected starting at the beginning of the gradient at 33 ml. **b** Gel filtration purification step showing elution with 2 ml fractions collected after the void volume of 42 ml. **c** ResourceQ purification step showing loading and elution with 2.5 ml fractions collected starting at the beginning of the gradient at 22 ml. For all traces, green highlighted regions indicate volumes taken for subsequent purification steps or as pure protein.

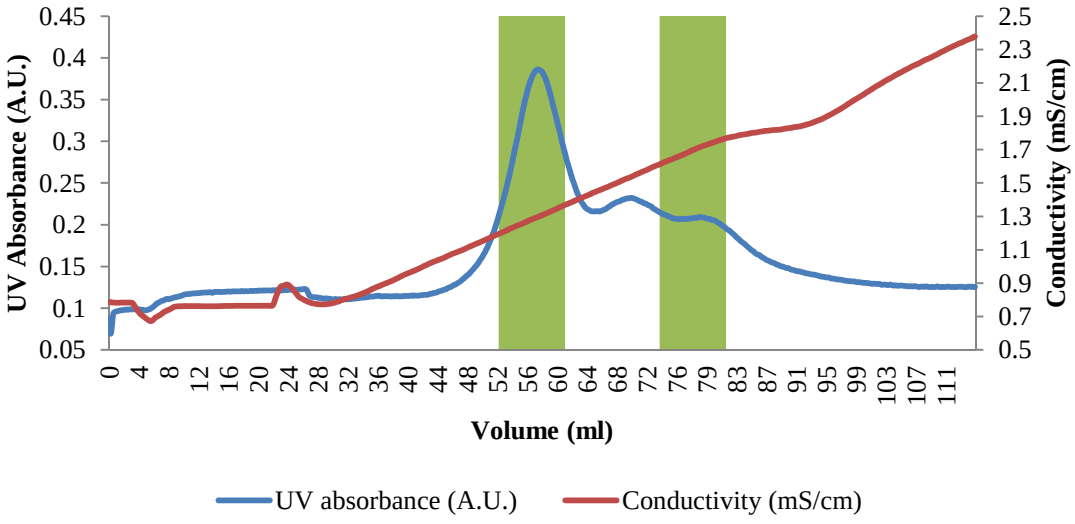
(a) BPSL0599 DEAE purification step



(b) BPSL0599 Superdex 200 purification step



(c) BPSL0599 ResourceQ purification step



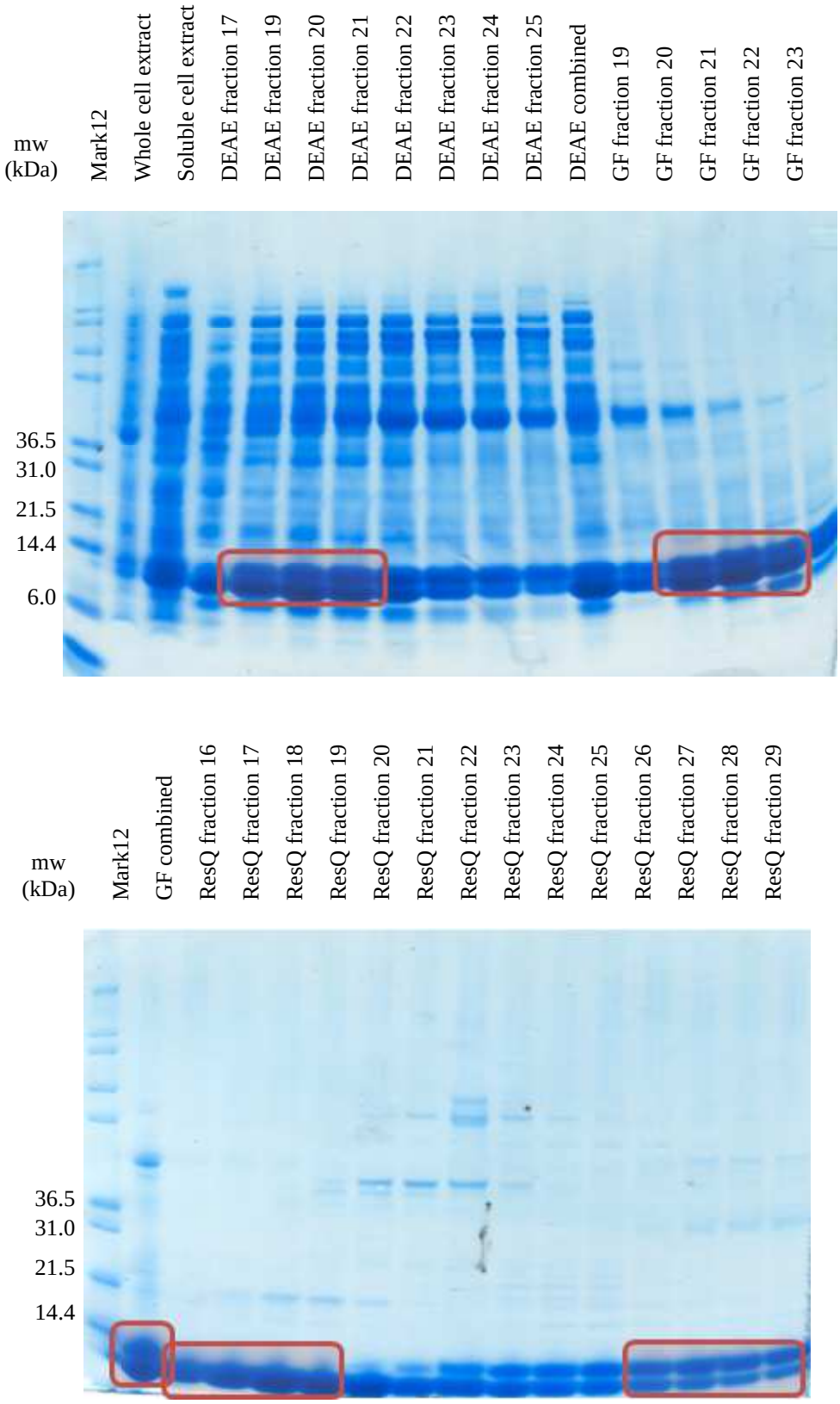


Figure 5.2 SDS-PAGE gels showing the purification of BPSL0599. The theoretical molecular weight of BPSL0599 is 13.8 kDa but it appears in two forms. The highlighted bands indicate the protein in fractions taken for subsequent purification steps or as one of the three protein samples used in crystallisation trials.

5.1.1 Purification analysis

The elution profile from the gel filtration column shows a single broad peak roughly corresponding to a trimeric form of the protein (figure 5.1 b). The mixed sample was sent for mass spectrometry (Simon Thorpe, University of Sheffield) to identify the two components. The larger protein is the full length BPSL0599 without its N-terminal methionine with the smaller protein representing a clipped fragment of BPSL0599 (figure 5.3). SDS-PAGE analysis of the purified samples shows they were of differing levels of purity (figure 5.2). The mixed sample contained approximately 45 % of each with 10 % other contaminants. The pure fragment sample contained only the fragment and approximately 5 % other contaminants. The full length sample was about 60 % pure with 35 % of the smaller fragment and 5 % other contaminants.

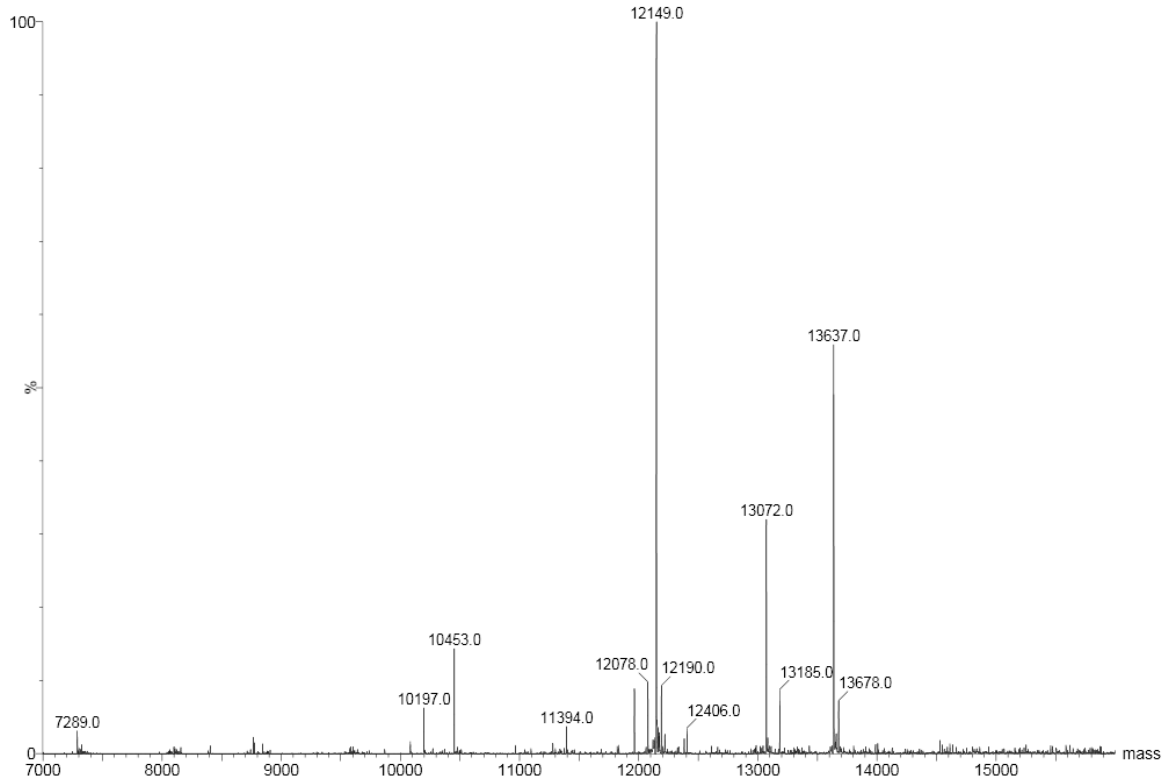
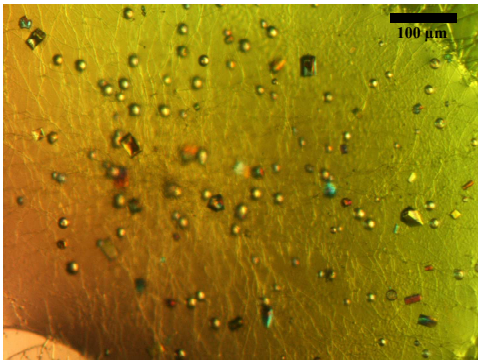


Figure 5.3 Mass spectrometry results for the mixed sample of purified BPSL0599. The molecular weight of 13637.0 Da represents the full length protein without its N-terminal methionine (residues 2 – 120) with the 12149.0 Da molecule being a clipped form of the protein (residues 2 – 106 or 3 – 107).

5.2 Protein crystallisation for BPSL0599

Five initial 96 condition robot screens, the $AmSO_4$, JCSG+, PACT, Pegs and Classics suites, were conducted using the three samples of purified protein (a mixed sample, and two samples predominantly containing the two different molecular weight proteins). 200 nl of protein was mixed with 200 nl of well solution and the trays were incubated at 17 ºC. Initial hits were obtained for the two single species samples, in condition Pegs G4 containing 200 mM calcium acetate and 20 % (w/v) PEG 3350 and for the full length sample in condition JCSG G11 containing 100 mM BIS-TRIS pH 5.5 and 2 M ammonium sulphate (figure 5.4). The initial hits are yet to be tested for X-ray diffraction or optimised to produce better crystals.

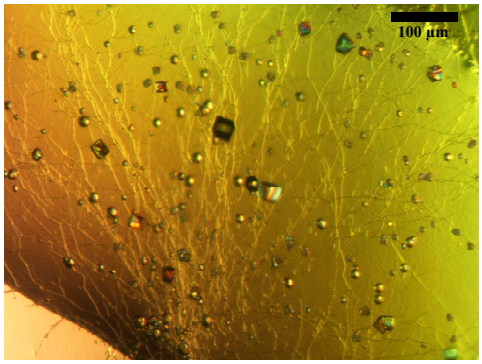


(2-120) – Pegs G4 – Robot screen

200 mM Calcium acetate

20 % (w/v) PEG 3350

Drop size 200 nl protein + 200 nl well solution

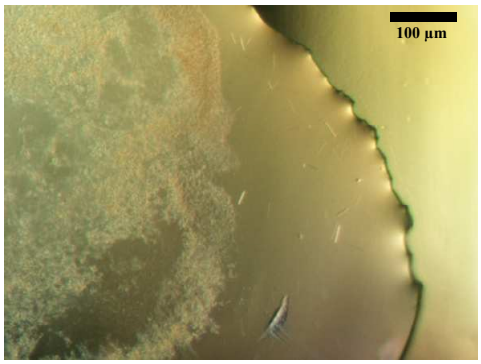


(2-106) – Pegs G4 – Robot screen

200 mM Calcium acetate

20 % (w/v) PEG 3350

Drop size 200 nl protein + 200 nl well solution



(2-120) – JCSG G11 – Robot screen

100 mM BIS-TRIS pH 5.5

2 M Ammonium sulphate

Drop size 200 nl protein + 200 nl well solution

Figure 5.4 Photographs of BPSL0599 crystals.

6.0 Studies on the protein BPSL1958

This section describes the purification, crystallisation, data collection and attempts to solve the structure for the target BPSL1958.

6.1 Protein purification for BPSL1958

6.1.1 Native protein purification

Approximately 3 g of cell paste was resuspended in 30 ml 25 mM TRIS pH 9.0 and disrupted by sonication. Cell debris and insoluble protein were removed by centrifugation at 70,000 g for 15 minutes and the supernatant was loaded onto a DEAE-Sepharose fast flow column equilibrated with 25 mM TRIS pH 9.0. A 200 ml gradient from 0 to 500 mM NaCl was then applied to the column and 8 ml fractions were collected (figure 6.1 a). Fractions were analysed by SDS-PAGE (figure 6.2) and BPSL1958 was found to interact weakly with the column eluting at the end of the loading step. Fractions containing the protein were pooled and concentrated to 1.5 ml using a Vivaspin concentrator with a 10 kDa MWCO. This was then loaded onto a 1.6 x 60 cm Superose 6 gel filtration column equilibrated in 50 mM TRIS pH 8.0 and 500 mM NaCl and eluted using the same buffer collecting 2 ml fractions (figure 5.1 b). Peak fractions were combined, the buffer was exchanged for 10 mM TRIS pH 8.0 and the protein was concentrated to 8 mg ml^{-1} for use in crystallisation trials using the Vivaspin concentrator. The overall yield of protein was low with only 0.7 mg being obtained which was estimated by SDS-PAGE to be over 95 % pure (figure 6.2).

6.1.2 Mutant protein purification

The mutant BPSL1958 proteins, K3C, S128C, A357C, K3C-D244C and K3C-H340C, created in order to allow phasing experiments for structure determination (sections 3.4 and 4.1.1), were purified using the same techniques as the native protein. Between 4 and 8 g of cell paste were used for each purification and the overall yields of mutant proteins were low, typically less than 2 mg in total. However BPSL1958 mutants containing the K3C mutation eluted from the DEAE column later than the native protein, suggesting slightly tighter binding, with most of the protein not eluting before the start of the NaCl gradient (figure 6.3). This is likely due to the loss of a positive charge on the proteins surface represented by the lysine to cysteine mutation.

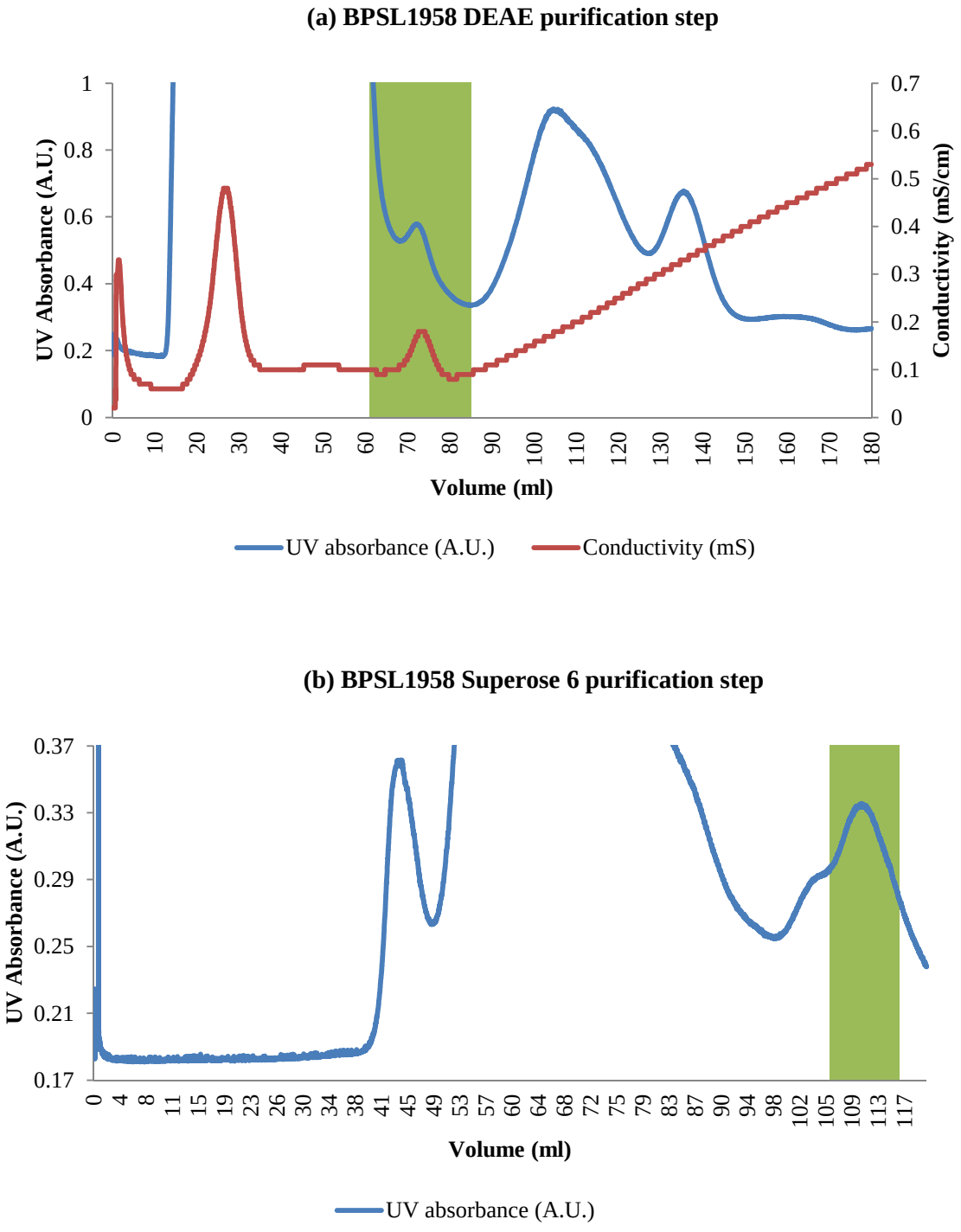


Figure 6.1 Chromatogram traces for the purification of BPSL1958. **a** DEAE purification step showing column loading and elution. 8 ml fractions were collected starting at the beginning of the gradient at 45 ml. **b** Gel filtration purification step showing elution with 2 ml fractions collected throughout. For both traces, highlighted regions indicate the eluant taken for subsequent purification step or as pure protein.

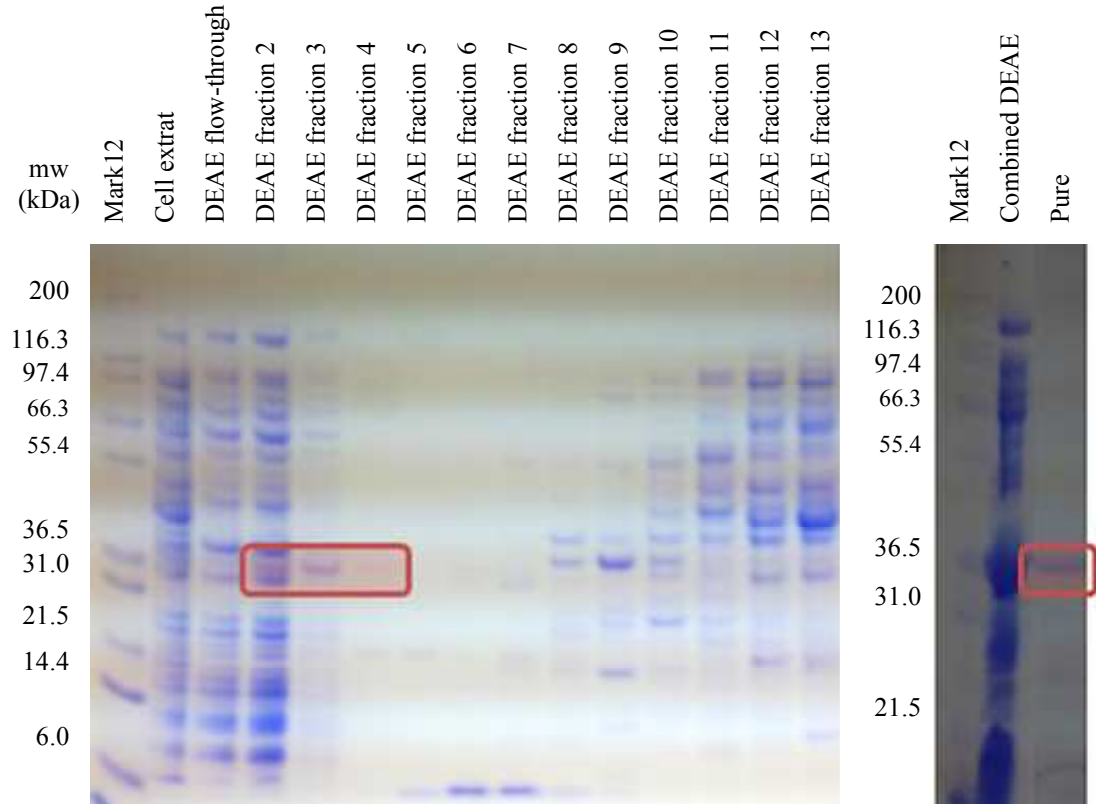


Figure 6.2 SDS-PAGE gel showing the purification of BPSL1958. The molecular weight of BPSL1958 is 36.1 kDa and the highlighted bands indicate the protein in fractions taken for subsequent purification steps or as pure protein.

6.1.3 Seleno-methionine mutant protein purification

The BPSL1958 K3C-I44M-I252M-I356M quadruple mutant seleno-methionine protein was purified using the same techniques as the other K3C mutant proteins. The overall yield was very low with only 0.5 mg being obtained which was concentrated to 8 mg ml^{-1} for use in crystallisation trials.

6.1.4 Purification analysis

The purified native protein was analysed by N-terminal sequencing (Arthur Moir, University of Sheffield) confirming it was BPSL1958 without its N-terminal methionine. During the purification the protein was found to elute significantly later than expected from the Superose 6 gel filtration column (figure 6.1 b) with a predicted molecular weight of approximately 10 kDa, not the expected 36.1 kDa. This suggests the protein interacts with the matrix of the column causing it to elute anomalously.

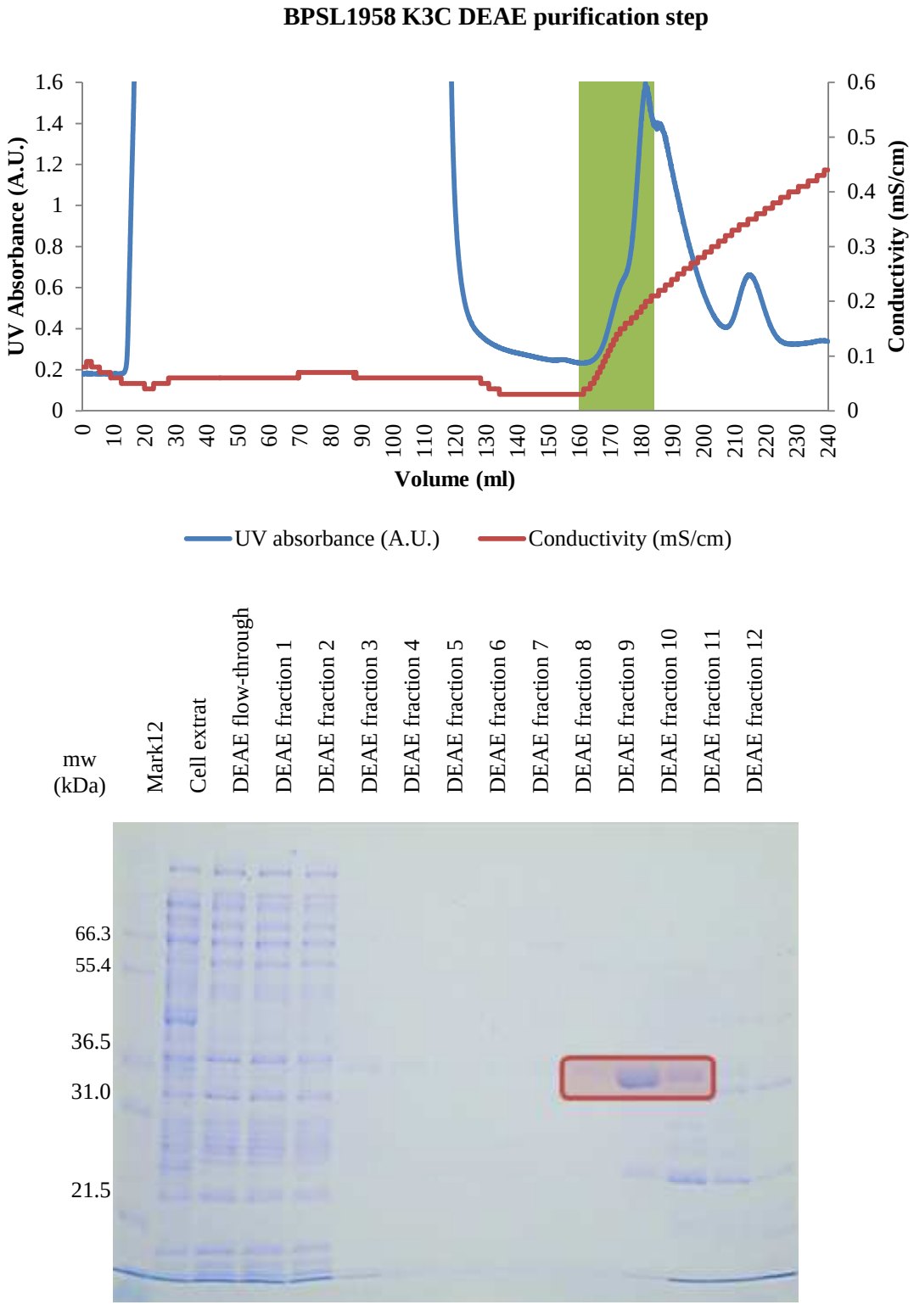
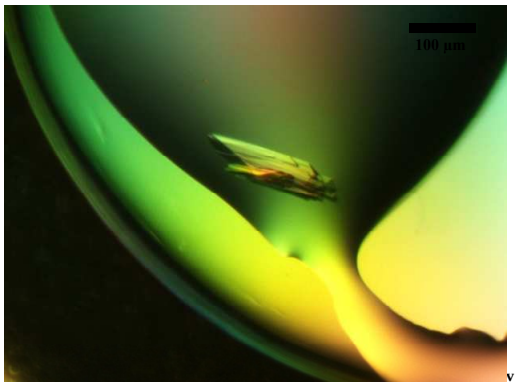


Figure 6.3 Chromatogram trace and SDS-PAGE analysis of BPSL1958 K3C DEAE purification. The trace shows loading and elution with 8 ml fractions collected starting at the beginning of the gradient at 104 ml. The highlighted region indicates the volume taken for the subsequent purification step. The SDS-PAGE gel shows fractions from the DEAE purification step of BPSL1958 K3C. The highlighted region indicates the fractions taken for subsequent purification steps.

6.2 Protein crystallisation for BPSL1958

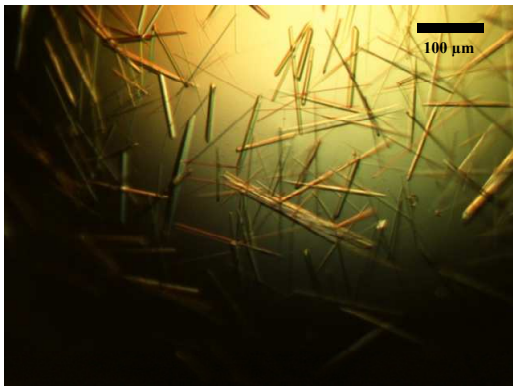
6.2.1 Native protein crystallisation

Two initial 96 condition screens, the Classics and Pegs suites, were conducted using purified BPSL1958 at 8 mg ml^{-1}. 200 nl of protein was mixed with 200 nl of well solution and the trays were incubated at 17 ºC. A single hit was found in the Classics robot crystal screen in well H11 (figure 6.4) containing 100 mM HEPES pH 7.5, 20 % (w/v) PEG 10,000 and 8 % (v/v) ethylene glycol. This condition was repeated in a single hanging drop experiment using a 5 µl protein, 5 µl well solution drop to obtain larger single crystals for data collection.



Classics H11 – Robot screen

100 mM HEPES pH 7.5
20 % (w/v) PEG 10000
8 % (v/v) Ethylene glycol
Drop size 200 nl protein + 200 nl well solution



Classics H11 – Hanging drop

100 mM HEPES pH 7.5
20 % (w/v) PEG 10000
8 % (v/v) Ethylene glycol
Drop size 5 µl protein + 5 µl well solution

Figure 6.4 Photographs of BPSL1958 crystals.

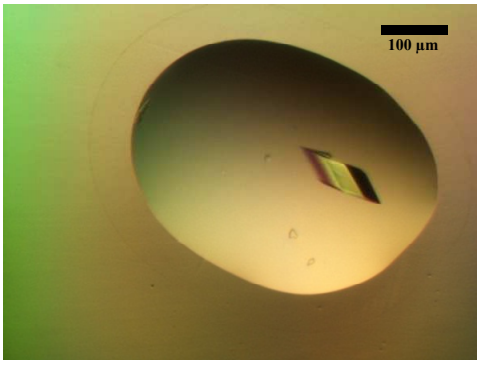
6.2.2 BPSL1958 K3C and S128C mutant protein crystallisation

Three initial 96 condition robot screens, the JCSG+, PACT and Pegs suites, were conducted using purified BPSL1958 K3C at 7 mg ml^{-1} and BPSL1958 S128C at 8 mg ml^{-1} on their own and mixed with a five times molar excess of EMTS. 200 nl of protein was mixed with 200 nl of well solution and the trays were incubated at 17 ºC. The only hit was found for BPSL1958 K3C in the Pegs robot crystal screen containing EMTS (figure 6.5) with no protein crystals being formed in any conditions for the S128C mutant protein. The single crystal was found in condition D8, 100 mM TRIS pH 8.5 and 25 % (w/v) PEG 4000. Attempts to optimise the crystals by altering the PEG concentration (20 – 35 % (w/v)) and pH (7.0 – 9.0) using 1 µl

protein solution and 1 µl well solution initially failed producing no crystals or precipitate. A further single experiment using the initial conditions and a drop size of 8 µl protein solution and 2 µl well solution produced several large crystals of the K3C mutant protein. Following this success, four further screens, the JCSG+, PACT, Pegs and Classics suites, were conducted for BPSL1958 S128C using protein at 8 mg ml^{-1} mixed with a ten times molar excess of EMTS. For these screens 500 nl of protein solution was mixed with 200 nl of well solution and the trays were incubated at 17 ºC. Several hits were found (figure 6.5) in conditions Pegs D5 containing 100 mM HEPES pH 7.5 and 20 % (w/v) PEG 10,000, Classics A5 containing 100 mM HEPES pH 7.5, 10 % (v/v) isopropanol and 20 % (w/v) PEG 4,000, Classics H3 containing 100 mM tri-sodium citrate pH 5.6, 200 mM ammonium acetate and 30 % (w/v) PEG 4,000, Classics H4 containing 100 mM TRIS pH 8.5, 200 mM magnesium chloride and 30 % (w/v) PEG 4,000, and Classics H8 containing 100 mM MES pH 6.5, 200 mM ammonium acetate and 30 % (w/v) PEG 5,000 MME. Optimisation trials were conducted for all initial BPSL1958 S128C crystal hits by altering the pH and concentration of components. For condition Pegs D5, the pH (7.0 – 8.0) and PEG concentration (15 – 30 % (w/v)) were varied. For condition Classics A5, the pH (7.0 – 8.0), isopropanol concentration (0 – 20 % (v/v)) and PEG concentration (15 – 30 % (w/v)) were varied. For condition Classics H3, the pH (5.0 – 6.0), and PEG concentration (20 – 40 % (w/v)) were varied. For condition Classics H4, the pH (8.0 – 9.0), and PEG concentration (20 – 40 % (w/v)) were varied. For condition Classics H8, the pH (6.0 – 7.0), and PEG concentration (20 – 40 % (w/v)) were varied. Optimisation screens for conditions Pegs D5 and Classics A5 were successful with the other optimisation screens failing to produce crystals that appeared superior to the initial hits on inspection. The best crystals were found in the optimised condition 100 mM HEPES pH 7.5, 5 % (v/v) isopropanol and 12 % (w/v) PEG 4,000, and these crystals were selected for data collection.

6.2.3 BPSL1958 K3C-D244C, K3C-H340C and A357C mutant protein crystallisation

Four initial 96 condition robot screens, the JCSG+, PACT, Pegs and Classics suites, were conducted using purified BPSL1958 K3C-D244C, K3C-H340C and A357C all at 8 mg ml^{-1} mixed with a five times molar excess of EMTS. The screens were conducted using both a 200 nl of protein mixed with 200 nl of well solution and 500 nl of protein mixed with 200 nl of well solution drop size and the trays were incubated at 17 ºC. Hits were found in a number of conditions for the three mutants (figures 6.6). Three conditions familiar to all three mutants and also BPSL1958 S128C were Classics H3, H4 and H8. Crystals were also found for

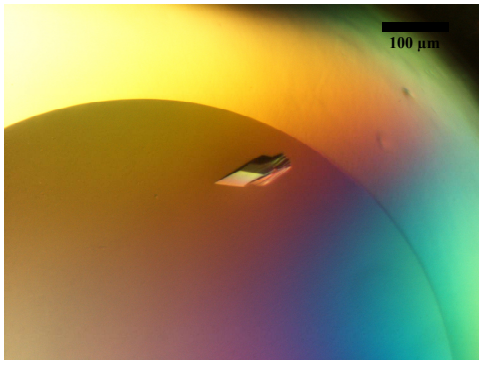


K3C – Pegs D8 – Robot screen

100 mM TRIS pH 8.5

25 % (w/v) PEG 4000

Drop size 200 nl protein + 200 nl well solution

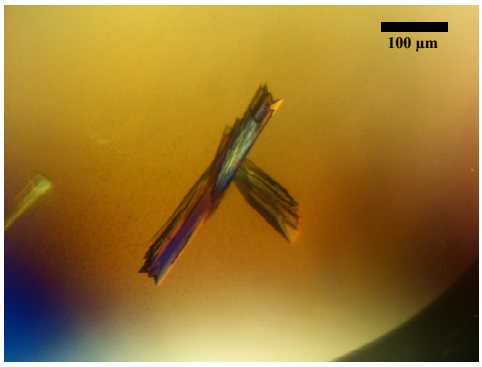


S128C – Pegs D5 – Robot screen

100 mM HEPES pH 7.5

20 % (w/v) PEG 10000

Drop size 500 nl protein + 200 nl well solution



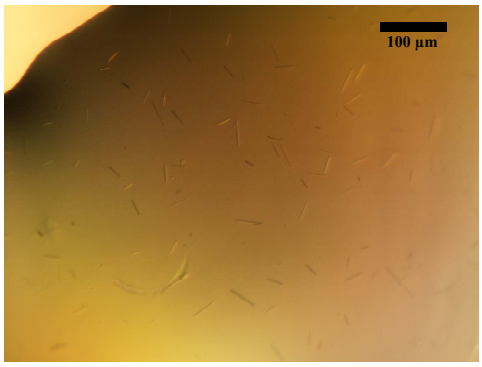
S128C – Classics A5 – Robot screen

100 mM HEPES pH 7.5

20 % (w/v) PEG 4000

10 % (v/v) Isopropanol

Drop size 500 nl protein + 200 nl well solution



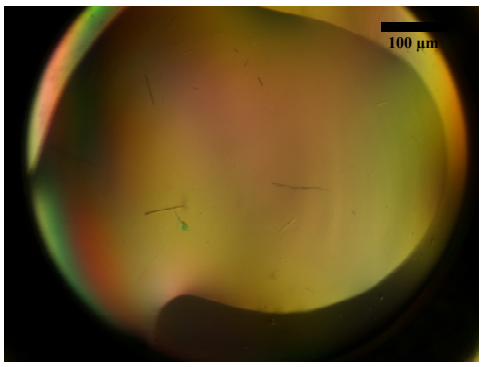
S128C – Classics H3 – Robot screen

200 mM Ammonium acetate

100 mM Tri-sodium citrate pH 5.6

30 % (w/v) PEG 4000

Drop size 500 nl protein + 200 nl well solution



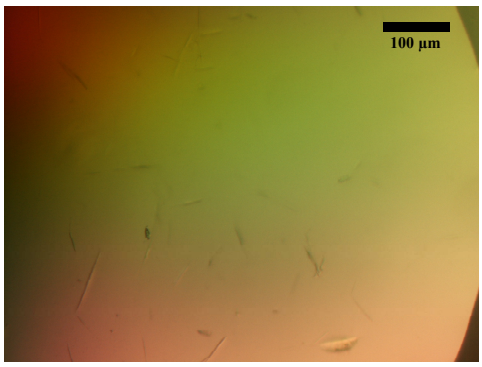
S128C – Classics H4 – Robot screen

200 mM Magnesium cholride

100 mM TRIS pH 8.5

30 % (w/v) PEG 4000

Drop size 500 nl protein + 200 nl well solution



S128C – Classics H8 – Robot screen

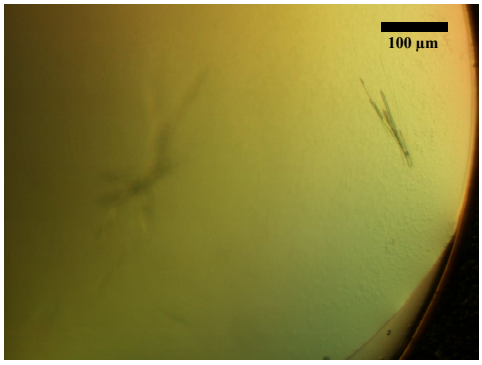
200 mM Ammonium sulphate

100 mM MES pH 6.5

30 % (w/v) PEG 5000 MME

Drop size 500 nl protein + 200 nl well solution

Figure 6.5 Photographs of BPSL1958 K3C + EMTS and BPSL1958 S128C + EMTS crystals.



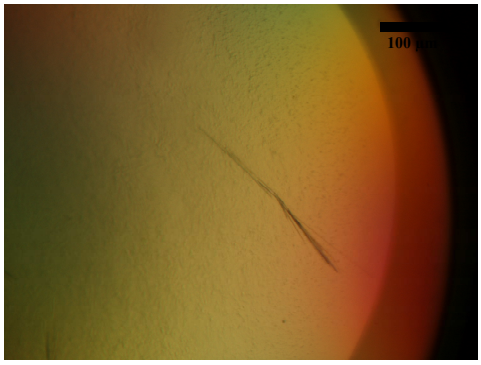
K3C-D244C – Classics H3 – Robot screen

200 mM Ammonium acetate

100 mM Tri-sodium citrate pH 5.6

30 % (w/v) PEG 4000

Drop size 500 nl protein + 200 nl well solution



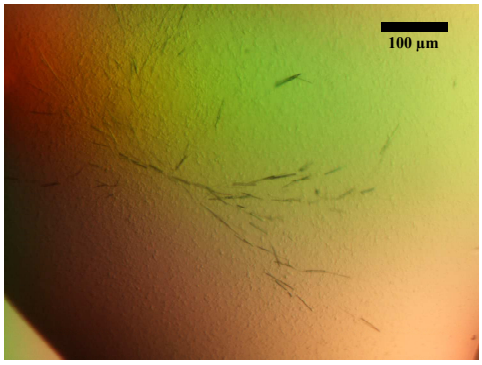
K3C-D244C – Classics H4 – Robot screen

200 mM Magnesium cholride

100 mM TRIS pH 8.5

30 % (w/v) PEG 4000

Drop size 500 nl protein + 200 nl well solution



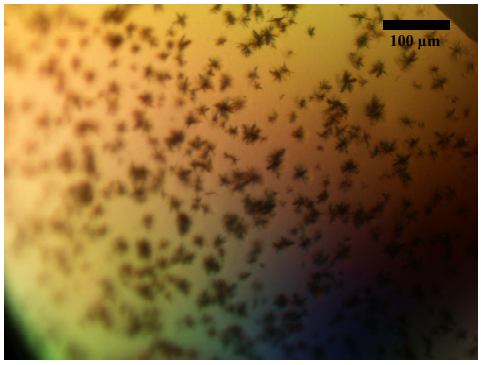
K3C-D244C – Classics H8 – Robot screen

200 mM Ammonium sulphate

100 mM MES pH 6.5

30 % (w/v) PEG 5000 MME

Drop size 500 nl protein + 200 nl well solution

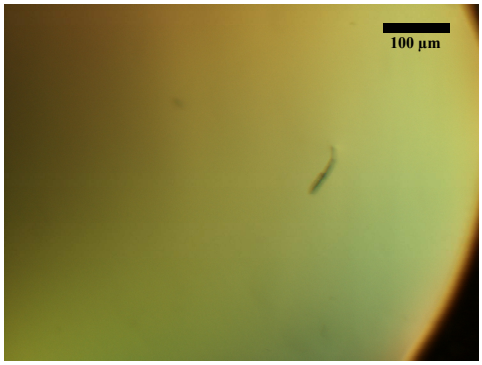


K3C-H340C – Classics D8 – Robot screen

100 mM HEPES pH 7.5

1.4 M Tri-sodium citrate

Drop size 500 nl protein + 200 nl well solution



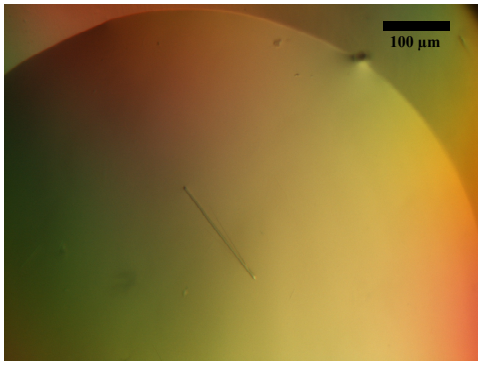
K3C-H340C – Classics H3 – Robot screen

200 mM Ammonium acetate

100 mM Tri-sodium citrate pH 5.6

30 % (w/v) PEG 4000

Drop size 500 nl protein + 200 nl well solution



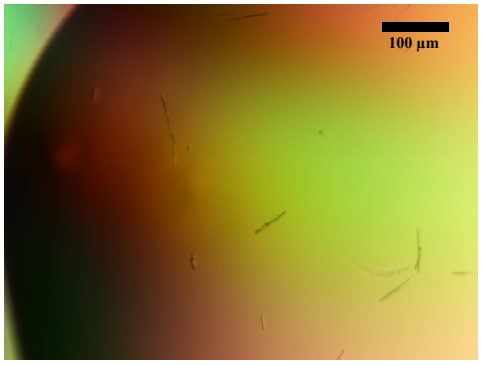
K3C-H340C – Classics H4 – Robot screen

200 mM Magnesium cholride

100 mM TRIS pH 8.5

30 % (w/v) PEG 4000

Drop size 500 nl protein + 200 nl well solution



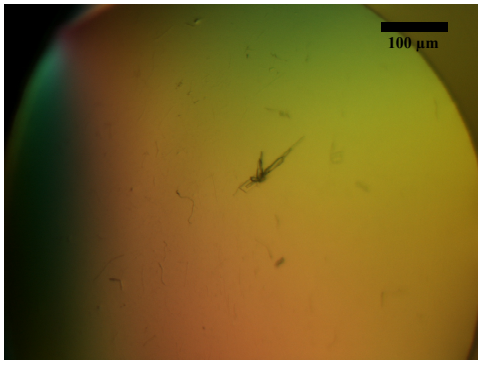
K3C-H340C – Classics H8 – Robot screen

200 mM Ammonium sulphate

100 mM MES pH 6.5

30 % (w/v) PEG 5000 MME

Drop size 500 nl protein + 200 nl well solution

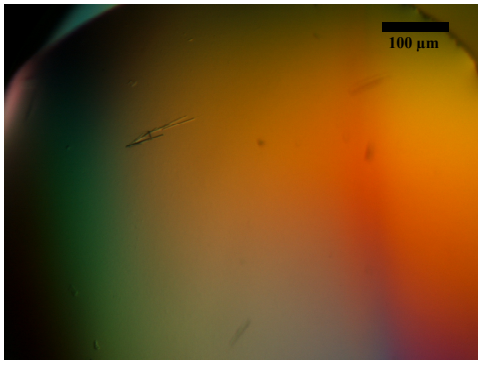


A357C – Classics D8 – Robot screen

100 mM HEPES pH 7.5

1.4 M tri-sodium citrate

Drop size 500 nl protein + 200 nl well solution



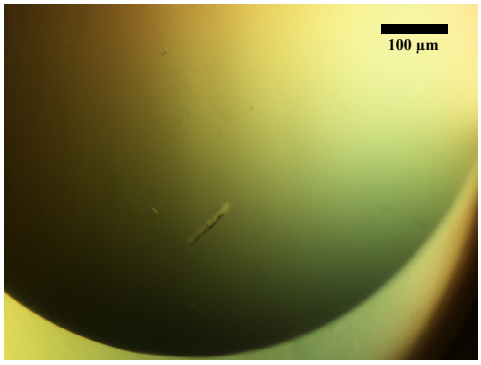
A357C – Classics F11 – Robot screen

200 mM Sodium acetate

100 mM Sodium cacodylate pH 6.5

30 % (w/v) PEG 8000

Drop size 500 nl protein + 200 nl well solution



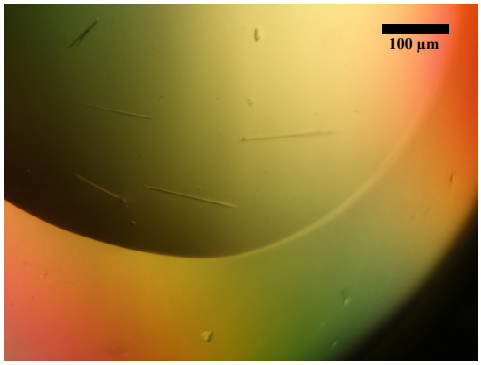
A357C – Classics H3 – Robot screen

200 mM Ammonium acetate

100 mM tri-sodium citrate pH 5.6

30 % (w/v) PEG 4000

Drop size 500 nl protein + 200 nl well solution



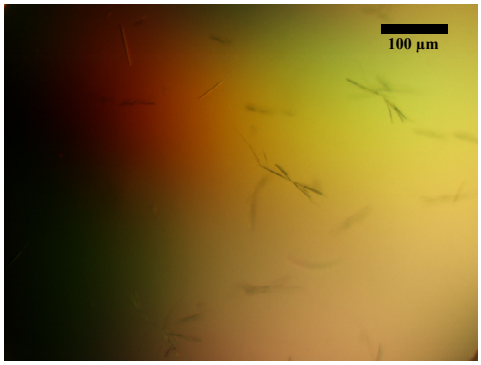
A357C – Classics H4 – Robot screen

200 mM Magnesium cholride

100 mM TRIS pH 8.5

30 % (w/v) PEG 4000

Drop size 500 nl protein + 200 nl well solution



A357C – Classics H8 – Robot screen

200 mM Ammonium sulphate

100 mM MES pH 6.5

30 % (w/v) PEG 5000 MME

Drop size 500 nl protein + 200 nl well solution

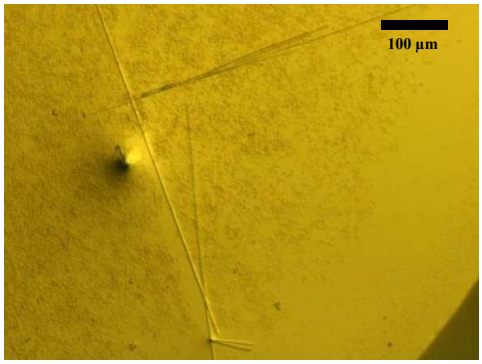
Figure 6.6 Photographs of BPSL1958 K3C-D244C + EMTS, BPSL1958 K3C-H340C and BPSL1958 A357C + EMTS crystals.

BPSL1958 K3CH340C and A357C in condition Classics D8 containing 100 mM HEPES pH 7.5 and 1.4 M tri-sodium citrate with a further two hits for BPSL1958 A357C in conditions Classics F11 containing 200 mM sodium acetate, 100 mM sodium cacodylate pH 6.5 and 30 % (w/v) PEG 8,000, and Classics H6 containing 200 mM sodium acetate, 100 mM TRIS pH 8.5 and 30 % (w/v) PEG 4,000. As many of the initial hit conditions were overlapping between the different mutants, optimisation screens were conducted around all hit conditions, found for any of the mutants or the native protein (Pegs A5 and D8, Classics A5, D8, F11, H3, H4, H6, H8 and H11), for BPSL1958 K3C-D244C, K3C-H340C and A357C. These were conducted, as for the other mutants, by varying the pH and concentration of the components. For condition Classics D8, the pH (7.0 – 8.0), and tri-sodium citrate concentration (1.0 – 2.0 M) were varied. For condition Classics F11, the pH (6.0 – 7.0), and PEG concentration (20 – 40 % (w/v)) were varied. For condition Classics H6, the pH (7.0 – 8.0), and PEG concentration (20 – 40 % (w/v)) were varied. For condition Classics H11, the pH (7.0 – 8.0), ethylene glycol concentration (0 – 15 % (v/v)) and PEG concentration (10 – 30 % (w/v)) were varied. Unfortunately attempts to optimise crystallisation conditions for the three mutants produced either no crystals or crystals of a similar quality to the robot trials. The most successful initial hits for BPSL1958 K3C-H340C and BPSL1958 A357C were Classics H4 and Classics F11 respectively, and crystals from these conditions were selected for data collection.

6.2.4 BPSL1958 K3C-I44M-I64M-I356M Seleno-methionine protein crystallisation

An initial screen around the successful crystallisation condition for the K3C single mutant, Pegs D8 with EMTS, by varying PEG concentration (20 – 30 % (w/v)) produced no crystals. Three forms of seeding were also attempted for the seleno-methionine protein using crystals of BPSL1958 K3C EMTS as seeds and the initial hit condition, Pegs D8 with EMTS. Macro-seeding was conducted by removing a crystal, briefly washing it in the well solution and placing into a drop containing 8 µl of seleno-methionine protein solution and 2 µl of well solution and suspended above wells containing crystallisation solution. Micro-seeding was attempted by looping crystals which were placed into a 10 µl drop of crystallisation solution on a clean coverslip. The crystals were ground up using a glass rod until they could no longer be seen with a microscope. The drop was then serially diluted to form six 90 µl solutions that were diluted 10^1 – 10^6 times from the original drop. 1 µl of these solutions were added to drops containing 7 µl of seleno-methionine protein solution and 2 µl of well solution and suspended above wells containing crystallisation solution. The final method was streak-

seeding which was attempted in two forms. The first was by using the 10^2 diluted solution from micro-seeding trials. A hair was dipped into the solution before being drawn across a drop containing 8 µl of seleno-methionine protein solution and 2 µl of well solution that had been allowed to equilibrate for five days and resuspended above wells containing crystallisation solution. The second method used a hair to rub against the edge of a formed crystal which was then drawn across a drop similarly to the first method. Some success was obtained through the streaking experiments using the micro-seeding solution, although crystals did not appear large enough for data collection (figure 6.7).



Pegs D8 – Streak-seeded

100 mM TRIS pH 8.5

25 % (w/v) PEG 4000

Drop size 8 µl protein + 2 µl well solution

Figure 6.7 Crystals of seleno-methionine BPSL1958 K3C-I44M-I64M-I356M + EMTS.

6.3 BPSL1958 native data

6.3.1 Native data collection

Native crystals were selected for data collection based on their size and overall definition from the successful optimisation trial. Crystals were looped and placed into a cryoprotectant solution consisting of the well solution with an increased concentration of 30 % (v/v) ethylene glycol. They were then looped again and flash frozen in a stream of gaseous nitrogen at 100 ºK and mounted onto the detector. Initial diffraction analysis was conducted in order to determine the diffraction quality of several crystals using an in-house Rigaku MM007 rotating anode generator and a MAR 345 research image plate. Two images were collected 90º apart with 1º oscillation. A number of crystals that diffracted beyond 2.5 Å on

the home source were saved and taken to a synchrotron at the I02 beamline of the Diamond light source, Oxford. Two initial test images were taken 90º apart with 1º oscillation to ensure crystal centering and to obtain a collection strategy using the auto-indexing and collection strategy components of Mosflm [146]. For the data set 420 images were collected with 0.25º oscillation per image using X-rays of 12658 eV at a crystal to detector distance of 255 mm using an ADSC Q315r detector. Data extending to 1.7 Å were collected (figure 6.8).

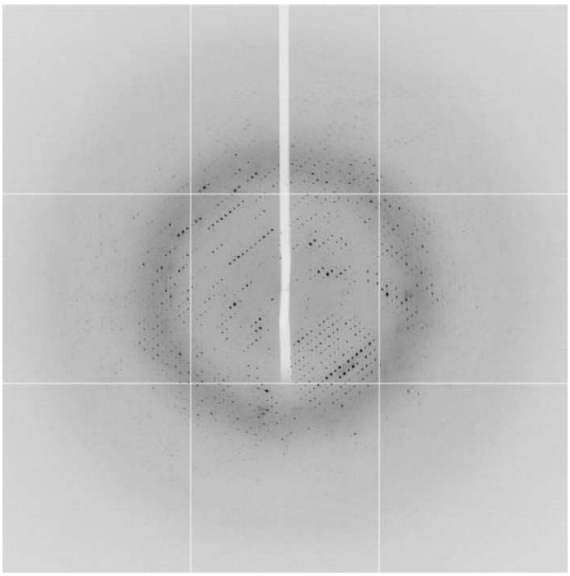


Figure 6.8 Diffraction image of the native crystal of BPSL1958.

6.3.2 Native data processing

The X-ray diffraction datasets were auto-processed at Diamond through the xia2 system using the 3dii mode [147]. Data were indexed and integrated by XDS and scaled by XSCALE [148]. The data indexed to a primitive orthorhombic space group with the unit cell parameters a = 63.7 Å, b = 99.3 Å, c = 108.2 Å (table 6.1). The space group was designated as most likely being $P2_12_12_1$ based on the apparent systematic absence of odd axial reflections for all three axes. Matthews coefficients calculated using Mattprob [149] showed the asymmetric unit was predicted to contain one, two or three protein molecules, with two being overwhelmingly the most likely, giving a Matthews coefficient of 2.37 and a solvent content of 48 % (table 6.2).

Dataset	Native protein crystal
Space group	Orthorhombic P
Unit cell parameters	
a (Å)	63.7
b (Å)	99.3
c (Å)	108.2
Energy (eV)	12658
Resolution range (Å)	45.14 – 1.73
Unique reflections	70099 (15397)
R_{merge}	0.059 (0.680)
R_{pim}	0.040 (0.531)
Completeness (%)	97.1 (80.7)
Multiplicity	2.2 (1.9)
Mean (I)/σ(I)	13.7 (2.0)

Table 6.1 Data collection statistics for the native BPSL1958 crystal. Numbers in parentheses indicate values for the highest resolution shell.

Molecules in the AU	Probability (based data resolution)	Probability (all proteins in the pdb)	Vm ($Å^3$ / Da)	Solvent content (%)	Molecular weight (Da)
1	0.0029	0.0203	4.74	74.04	36110
2	0.9944	0.9753	2.37	48.07	72220
3	0.0027	0.0045	1.58	22.11	108330

Table 6.2 Matthews coefficient calculations and probabilities for native crystals of BPSL1958. The results show a possibility of one, two or three protein molecules inhabiting the asymmetric unit.

6.4 Phasing by molecular replacement for BPSL1958

The two threading models for the six or seven bladed β-propeller proteins output by the Phyre 2 server [141] (figure 3.6) were used to create possible search models for molecular replacement using the native diffraction data. The two models were cut back to a poly alanine chain using chainsaw [150] before use for an automated search in Phaser [151]. The data were input and Phaser was run in $P2_12_12_1$ and all alternative related space groups. Unfortunately the best results for both search models were of a poor quality with low Z-scores for both the rotation and translation searches. The resulting refined models were also unconvincing after refinement with REFMAC5 [152] suggesting a solution had not been found (table 6.3).

	6-Bladed model	7-bladed model
Space group	$P222_1$	P222
Rotation function Z-score	3.8	4.8
Translation function Z-score	8.2	5.9
R	61.7	62.0

Table 6.3 Best molecular replacement solutions for BPSL1958 using threaded homology models.

6.5 BPSL1958 mutant data

6.5.1 K3C mutant EMTS co-crystallisation data collection

The initial robot trial hit was selected for data collection. The crystal was looped and placed into a cryoprotectant solution consisting of the well solution with the addition of 30 % (v/v) ethylene glycol. They were then looped again and flash frozen in a stream of gaseous nitrogen at 100 ºK and mounted onto the detector. Initial diffraction analysis was conducted in order to determine the diffraction quality of the crystal using an in-house Rigaku MM007 rotating anode generator and a MAR 345 research image plate. Two images were collected 90º apart with 1º oscillation and the diffraction was found to be of good quality. The crystal was saved and data was collected using a synchrotron at the I04 beamline of the Diamond light source, Oxford. Two initial test images were taken 90º apart with 1º oscillation to ensure crystal centering and to obtain a collection strategy using the auto-indexing and collection strategy components of Mosflm [146]. A fluorescence scan at the mercury L_{III} edge was conducted to select X-ray energies at which to collect data in a MAD experiment (figure 6.9a). Four energies were selected based on the fluorescence absorbance spectrum profile. The peak (maximised f'') and inflection point (minimised f') were selected as 12287 eV and 12284 eV, and a high and low energy remote dataset were selected as 12260 eV and 12295 eV respectively. For each wavelength 720 images were collected with 0.5º oscillation per image at a crystal to detector distance of 414 mm using an ADSC Q315r detector. Data extending to approximately 2.1 Å were collected for each X-ray energy (figure 6.10).

6.5.2 S128C mutant EMTS co-crystallisation data collection

Crystals were selected for data collection based on their size and overall definition from the successful optimisation trials. Crystals were looped and placed into a cryoprotectant solution consisting of the well solution with the addition of 30 % (v/v) ethylene glycol. They were then looped again and flash frozen in a stream of gaseous nitrogen at 100 ºK and mounted

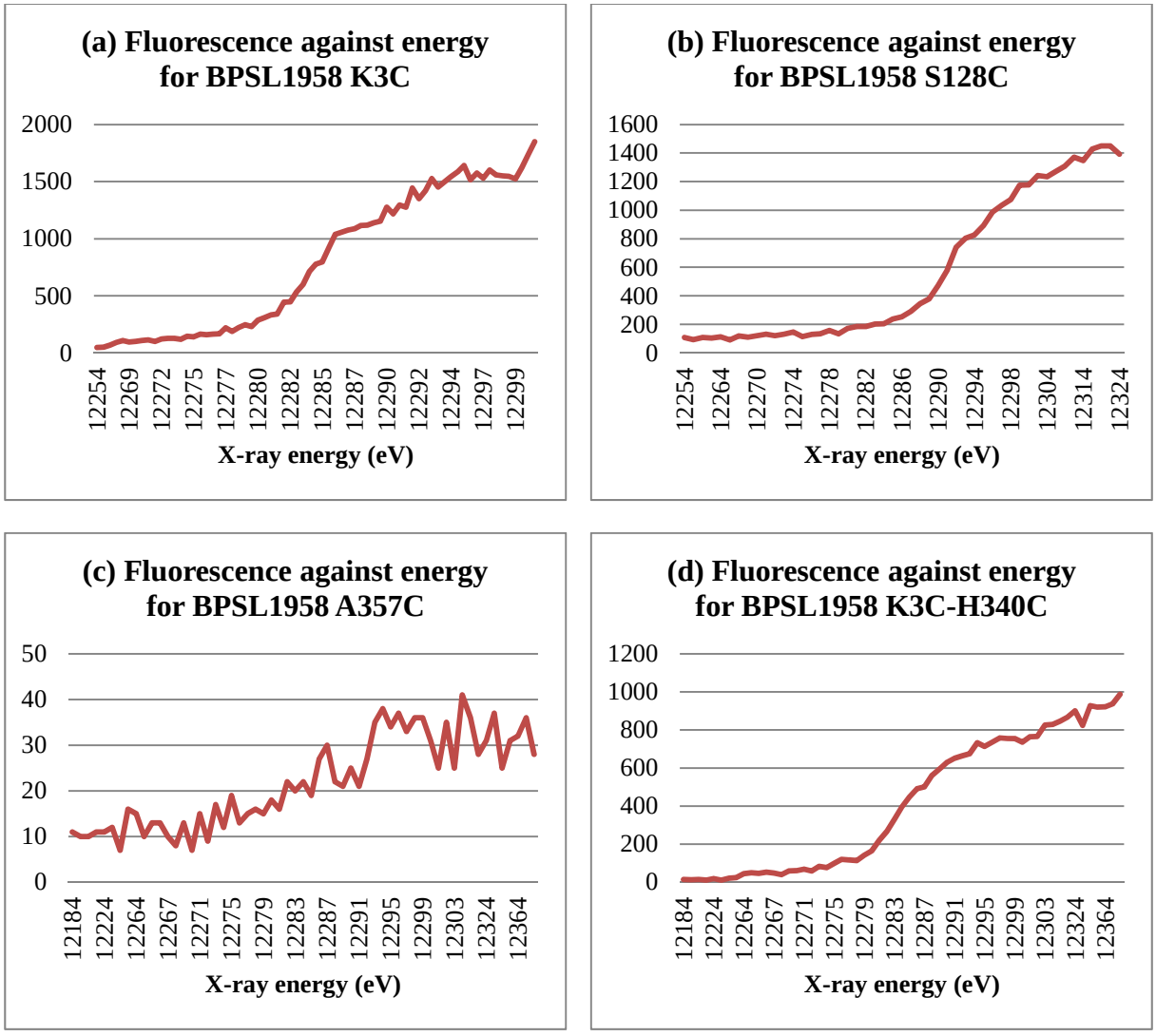
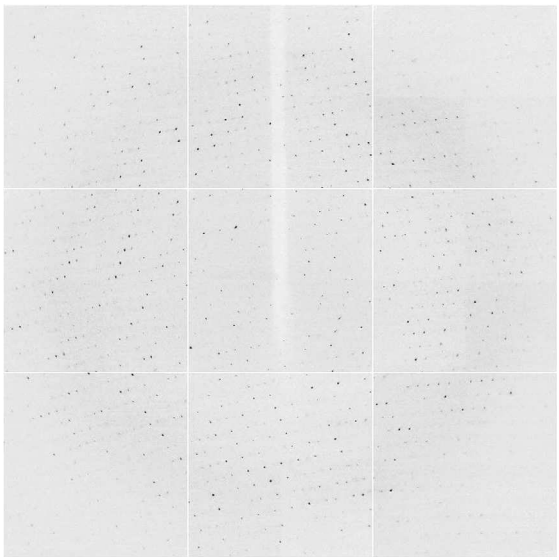
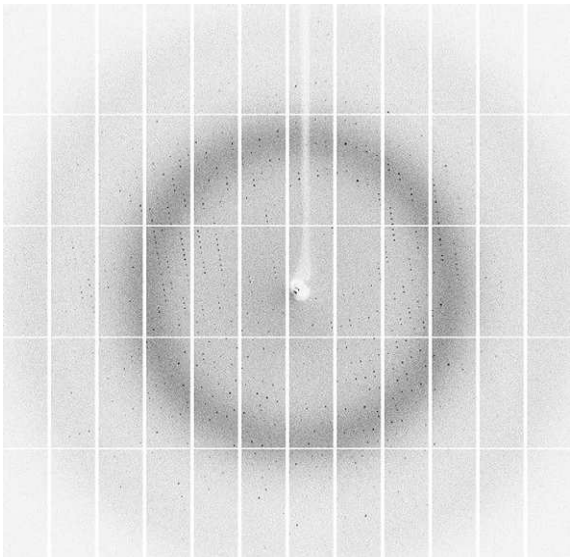


Figure 6.9 Mercury L_{III}-edge fluorescence scans for the BPSL1958 (EMTS) mutant crystals.

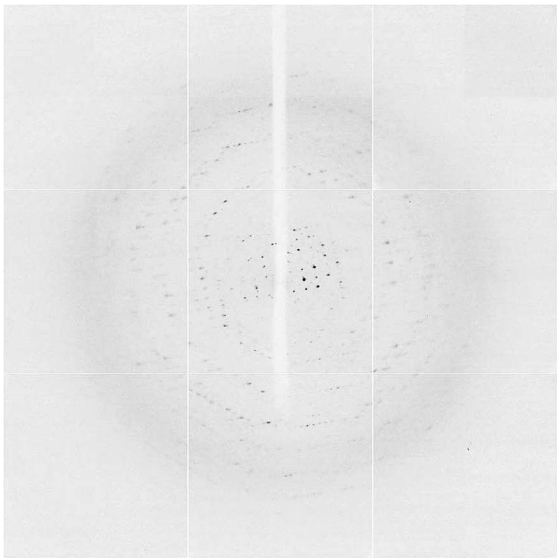
onto the detector. Several crystals were saved and data was collected using a synchrotron at the I24 beamline of the Diamond light source, Oxford. Two initial test images were taken 90º apart with 1º oscillation to ensure crystal centering and to obtain a collection strategy using the auto-indexing and collection strategy components of Mosflm [146]. A fluorescence scan at the mercury L_{III} edge was conducted to select X-ray energies at which to collect data in a MAD experiment (figure 6.9b). Three energies were selected based on the fluorescence absorbance spectrum profile. The peak (maximised f'') and inflection point (minimised f') were selected as 12294 eV and 12289 eV respectively and the high energy remote (maximised Δf' from inflection) was selected as 12500 eV. For each wavelength 450 images were collected with 0.2º oscillation per image at a crystal to detector distance of 400 mm using a Pilatus 6M detector. Data extending to approximately 2.0 Å were collected for each X-ray energy (figure 6.10).



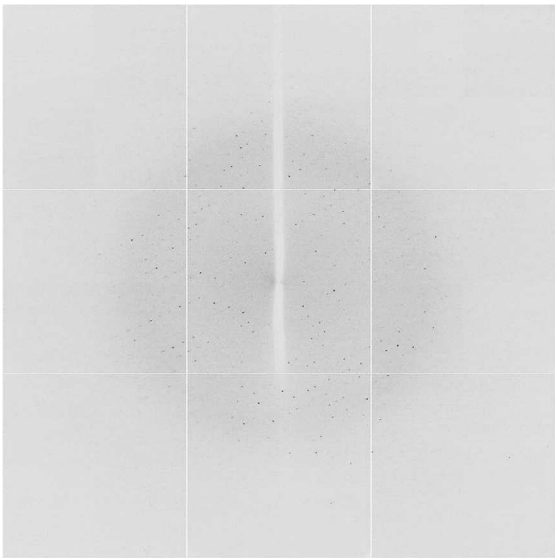
BPSL1958 K3C peak dataset



BPSL1958 S128C peak dataset



BPSL1958 A357C peak dataset



BPSL1958 K3C-H340C peak dataset

Figure 6.10 Diffraction images of crystals of four mutants of BPSL1958.

6.5.3 A357C mutant EMTS co-crystallisation data collection

Crystals were selected for data collection based on their size and overall definition from the initial robot trials. Crystals were looped and placed into a cryoprotectant solution consisting of the well solution with the addition of 30 % (v/v) ethylene glycol. They were then looped again and flash frozen in a stream of gaseous nitrogen at 100 ºK and mounted onto the detector. Several crystals were saved and data was collected using a synchrotron at the I04 beamline of the Diamond light source, Oxford. Two initial test images were taken 90º apart with 1º oscillation to ensure crystal centering and to obtain a collection strategy using the auto-indexing and collection strategy components of Mosflm [146]. A fluorescence scan at

the mercury L_{III} edge was conducted to select X-ray energies at which to collect data in a MAD experiment (figure 6.9c). The scan suggested the absence of mercury in the crystal as there was very little excitation, however two wavelengths were selected corresponding to 6 eV above the hypothetical peak at 12290 eV, and the hypothetical inflection at 12281 eV for a mercury derivative and used to collect full datasets. For both wavelengths 180 images were collected with 0.5º oscillation per image at a crystal to detector distance of 363 mm using an ADSC Q315r detector. Data extending to approximately 2.9 Å were collected for both X-ray energies (figure 6.10).

6.5.4 K3C-H340C mutant EMTS co-crystallisation data collection

Crystals were selected for data collection based on their size and overall definition from the initial robot trials. Crystals were looped and placed into a cryoprotectant solution consisting of the well solution with the addition of 30 % (v/v) ethylene glycol. They were then looped again and flash frozen in a stream of gaseous nitrogen at 100 ºK and mounted onto the detector. Several crystals were saved and data was collected using a synchrotron at the I04 beamline of the Diamond light source, Oxford. Two initial test images were taken 90º apart with 1º oscillation to ensure crystal centering and to obtain a collection strategy using the auto-indexing and collection strategy components of Mosflm [146]. A fluorescence scan at the mercury L_{III} edge was conducted to select X-ray energies at which to collect data in a MAD experiment (figure 6.9d). Three energies were selected based on the fluorescence absorbance spectrum profile. The peak (maximised f'') and inflection point (minimised f') were selected as 12291 eV and 12281 eV respectively, and high energy remote (maximised Δf' from inflection) was selected as 12400 eV. For each wavelength 360 images were collected with 0.5º oscillation per image at a crystal to detector distance of 280 mm using an ADSC Q315r detector. Data extending to approximately 2.9 Å were collected for each X-ray energy (figure 6.10).

6.5.5 Mutant data processing

The mutant X-ray diffraction datasets were auto-processed at Diamond through the xia2 system using the 3dii mode [147]. Data were indexed and integrated by XDS and scaled by XSCALE [148]. The K3C mutant datasets indexed to a primitive orthorhombic space group, most likely $P2_12_12_1$ based on systematic absences, with the unit cell parameters a = 63.2 Å, b = 99.0 Å, c = 109.2 Å, roughly corresponding to the native dataset (table 6.4). The K3C datasets contained data to approximately 2.1 Å however it was incomplete beyond

approximately 2.5 Å (figure 6.11) producing poor statistics for the overall data. The S128C datasets indexed to a primitive orthorhombic space group, most likely $P2_12_12_1$ based on systematic absences, with the unit cell parameters a = 63.5 Å, b = 99.0 Å, c = 108.5 Å, indicating that the crystals were isomorphous with the native dataset (table 6.5). The A357C datasets indexed to a primitive orthorhombic space group, with the unit cell parameters a = 63.6 Å, b = 99.4 Å, c = 107.9 Å. For this mutant, the programme Pointless [153] suggested the pattern of the space group was most likely to have a 2-fold axis rather than a 2-fold screw along a. Analysis of the raw data showed that this was based on the presence of a 12 σ reflection for the 5 0 0, however the intensity of the reflection is still much weaker than the vast majority of the even ordered reflections along the a-axis. Whether the assignment is correct, or an artefact arising from poor data is unclear, though the overall similarity between the unit cell dimensions and the pattern of reflections elsewhere would support the view that the crystal form is the same as for the native data. Therefore the data was treated as belonging to the same space group as the native, K3C and S128C datasets. The K3C-H340C datasets indexed to the space group $P2_1$, with the unit cell parameters a = 46.6 Å, b = 46.1 Å, c = 98.5 Å, β = 96.1°, showing a unique space group among the datasets (table 6.7). Matthews coefficients were approximately the same for the K3C, S128C and A357C datasets due to the similarity in unit cell parameters. The Matthews coefficient was calculated using Mattprob [149] for the K3C-H340C datasets and showed the asymmetric unit was predicted to contain one or two protein molecules, with one being overwhelmingly the most likely, giving a Matthews coefficients of 2.91 and a solvent content of 58 % (Table 6.8).

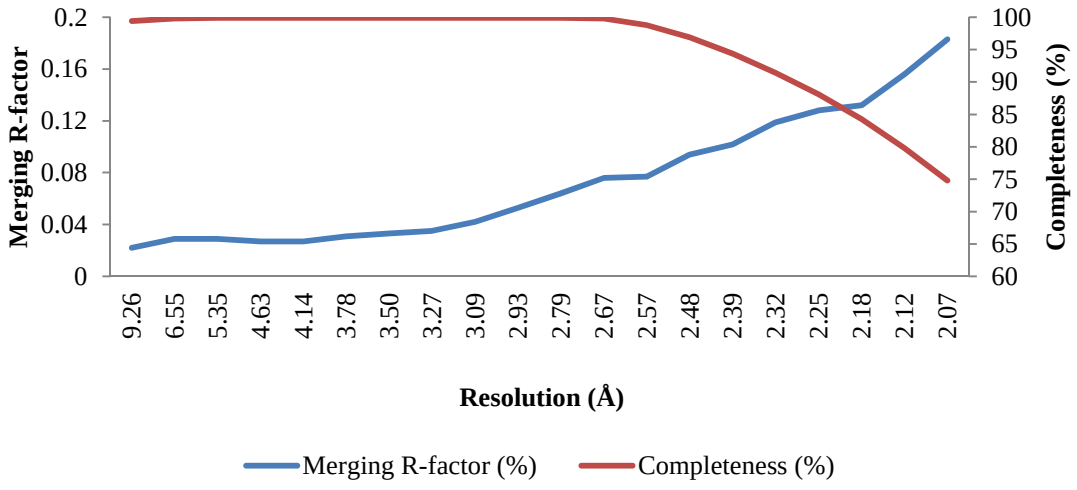


Figure 6.11 Data collection statistics against resolution for the BPSL1958 K3C data. The graph shows R_{merge} and completeness for the peak dataset. The graph suggests the data is complete and of a good quality to 2.5 Å.

	Low energy remote	Inflection	Peak	High energy remote
Space group	Orthorhombic P	Orthorhombic P	Orthorhombic P	Orthorhombic P
Unit cell parameters				
a (Å)	63.3	63.2	63.2	63.3
b (Å)	99.0	98.9	98.8	99.0
c (Å)	109.4	109.2	109.3	109.3
Energy (eV)	12260	12284	12287	12295
Resolution range (Å)	73.40 – 2.07	73.28 – 2.07	73.31 – 2.07	73.36 – 2.07
Unique reflections	31603 (328)	31522 (334)	31614 (352)	31608 (329)
R_{merge}	0.038 (0.233)	0.043 (0.249)	0.043 (0.183)	0.043 (0.222)
R_{pim}	0.012 (0.176)	0.012 (0.137)	0.012 (0.135)	0.014 (0.167)
Completeness (%)	74.4 (10.8)	74.5 (11.1)	74.8 (11.7)	74.4 (10.8)
Anomalous completeness (%)	72.5 (4.3)	72.7 (4.6)	72.9 (5.3)	72.6 (4.5)
Multiplicity	11.1 (2.4)	11.1 (2.5)	11.1 (2.6)	11.2 (2.5)
Anomalous multiplicity	5.9 (2.0)	5.9 (2.0)	5.9 (2.0)	5.9 (2.0)
Mean (I)/σ(I)	39.7 (3.4)	40.2 (4.4)	39.5 (4.5)	35.5 (3.3)

Table 6.4 Data collection statistics for the BPSL1958 K3C crystal. Numbers in parentheses indicate values for the highest resolution shell.

	Inflection	Peak	High energy remote
Space group	Orthorhombic P	Orthorhombic P	Orthorhombic P
Unit cell parameters			
a (Å)	63.5	63.5	63.6
b (Å)	99.3	98.9	99.2
c (Å)	108.0	108.5	108.2
Energy (eV)	12289	12294	12500
Resolution range (Å)	41.15 – 1.99	45.01 – 2.15	41.19 – 1.81
Unique reflections	46473 (3253)	37128 (2722)	57674 (2705)
R_{merge}	0.092 (0.386)	0.124 (0.598)	0.048 (0.427)
R_{pim}	0.049 (0.313)	0.089 (0.432)	0.039 (0.318)
Completeness (%)	98.0 (94.4)	98.3 (98.5)	92.2 (60.5)
Anomalous completeness (%)	91.3 (87.3)	92.0 (92.3)	82.0 (46.6)
Multiplicity	3.3 (3.1)	3.3 (3.3)	3.1 (2.5)
Anomalous multiplicity	1.8 (1.6)	1.8 (1.8)	1.7 (1.4)
Mean (I)/σ(I)	11.0 (2.3)	7.5 (2.2)	12.1 (2.1)

Table 6.5 Data collection statistics for the BPSL1958 S128C crystal. Numbers in parentheses indicate values for the highest resolution shell.

	Inflection	Peak
Space group	Orthorhombic P	Orthorhombic P
Unit cell parameters		
a (Å)	99.4	99.4
b (Å)	107.9	107.9
c (Å)	63.7	63.6
Energy (eV)	12281	12290
Resolution range (Å)	45.16 – 2.94	45.14 – 2.92
Unique reflections	14787 (1027)	15064 (1041)
R_{merge}	0.139 (0.750)	0.184 (0.815)
R_{pim}	0.097 (0.440)	0.097 (0.465)
Completeness (%)	97.9 (94.2)	97.8 (93.2)
Anomalous completeness (%)	89.7 (79.3)	89.2 (78.5)
Multiplicity	3.3 (3.1)	3.3 (3.1)
Anomalous multiplicity	1.9 (1.8)	1.9 (1.8)
Mean (I)/σ(I)	6.3 (1.9)	6.3 (1.8)

Table 6.6 Data collection statistics for the BPSL1958 A357C crystal. Numbers in parentheses indicate values for the highest resolution shell.

	Inflection	Peak	High energy remote
Space group	$P2_1$	$P2_1$	$P2_1$
Unit cell parameters			
a (Å)	46.5	46.6	46.6
b (Å)	46.1	46.1	46.2
c (Å)	98.2	98.5	98.5
β (º)	95.9	96.1	96.2
Energy (eV)	12281	12291	12400
Resolution range (Å)	48.86 – 2.82	97.92 – 2.92	48.99 – 2.67
Unique reflections	10151 (753)	9259 (666)	12043 (9.8)
R_{merge}	0.261 (0.682)	0.301 (0.717)	0.340 (0.868)
R_{pim}	0.190 (0.475)	0.130 (0.305)	0.177 (0.448)
Completeness (%)	99.2 (99.9)	99.9 (99.8)	99.5 (99.8)
Anomalous completeness (%)	90.9 (92.3)	99.7 (99.8)	92.8 (92.8)
Multiplicity	3.6 (3.7)	7.3 (7.5)	3.6 (3.7)
Anomalous multiplicity	1.9 (1.9)	3.8 (3.9)	1.9 (1.9)
Mean (I)/σ(I)	4.7 (1.8)	6.5 (2.6)	5.0(1.9)

Table 6.7 Data collection statistics for the BPSL1958 K3C-H340C crystal. Numbers in parentheses indicate values for the highest resolution shell.

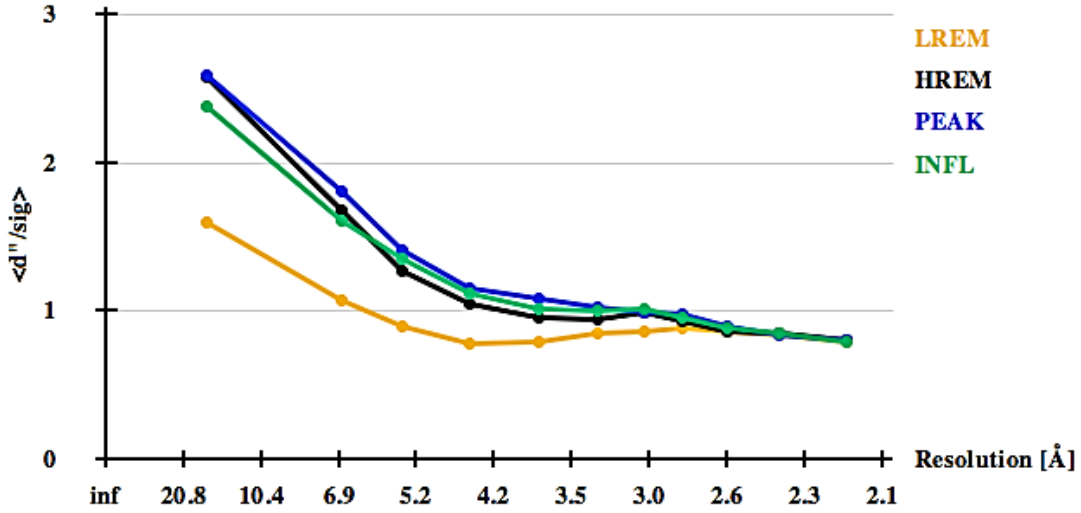
Molecules in the AU	Probability (based on resolution)	Probability (all proteins in the pdb)	Vm (Å^3 / Da)	Solvent content (%)	Molecular weight (Da)
1	0.9925	0.9918	2.91	57.78	36110
2	0.0075	0.0082	1.46	15.56	72220

Table 6.8 Matthews coefficient calculations and probabilities for crystals of BPSL1958 K3C-H340C. The results show a possibility of one or two protein molecules inhabiting the asymmetric unit.

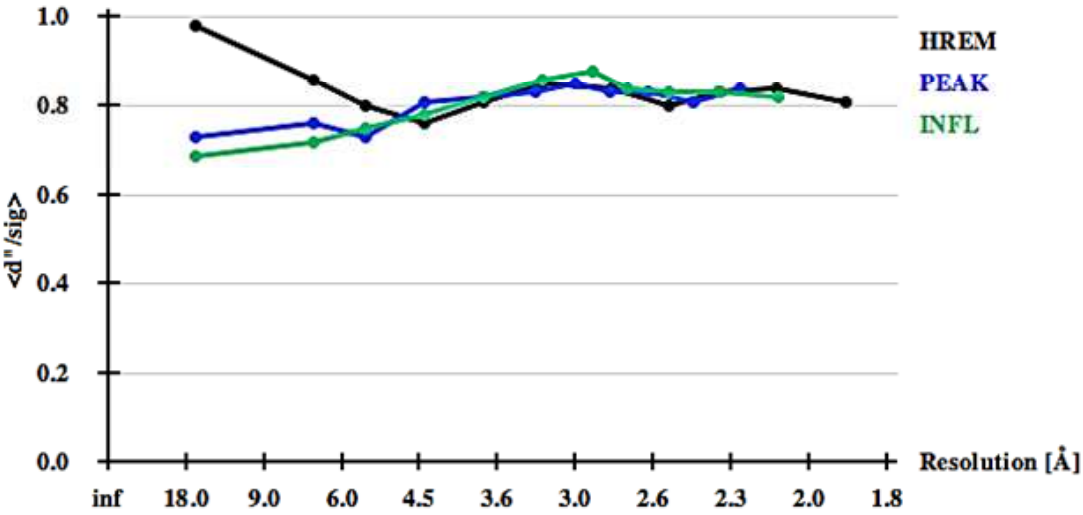
6.6 Experimental phasing for BPSL1958

Initial phasing was attempted with the programmes of the SHELX package [154] using the HKL2MAP graphical user interface for SHELXC and SHELXD [155]. SHELXC was first used to judge the presence and quality of any anomalous signal within the datasets from all the crystals. This suggested that the only datasets with any anomalous signal came from the BPSL1958 K3C crystal from the initial crystal trial, and this contained significant anomalous signal to 4.5 Å and usable signal to 3.2 Å (figure 6.12). SHELXD was used to calculate the heavy atom substructure in all possible primitive orthorhombic space groups using all data to 3.5 Å. The best solution was identified with a correlation coefficient of 51.13 and a Patterson figure of merit of 26.29. A total of six potential heavy atom sites were found, two very strong sites, two weaker sites and two very weak sites (figure 6.13). Preliminary protein phasing, density modification and initial model building were conducted using SHELXE-beta with auto-tracing. Five rounds of twenty cycles of phase calculation and density modification, including NCS averaging, followed by auto-tracing were carried out for both the original and the inverted hands of the mercury substructure. A solvent content of 48 % was used corresponding to two molecules in the asymmetric unit. The solutions for the two substructure hands gave values of 0.439 and 0.639, and 0.445 and 0.649, for the contrast and connectivity of the original and inverted hands respectively, preventing the assignment of a hand to the substructure based on these figures of merit. SHELXE produced models for the original and inverted hands consisting of 104 and 108 residues respectively, with correlation coefficients for partial structure against the data of 11.24 % and 7.29 %. The models for the two hands, if correct, only corresponded to a small fraction of the expected protein in the asymmetric unit and were unconvincing when viewed alongside the electron density maps. Furthermore inspection of the maps for the two hands provided no insight into the correct substructure hand, or provided a base for further model building, as both were equally uninterpretable (figure 6.14).

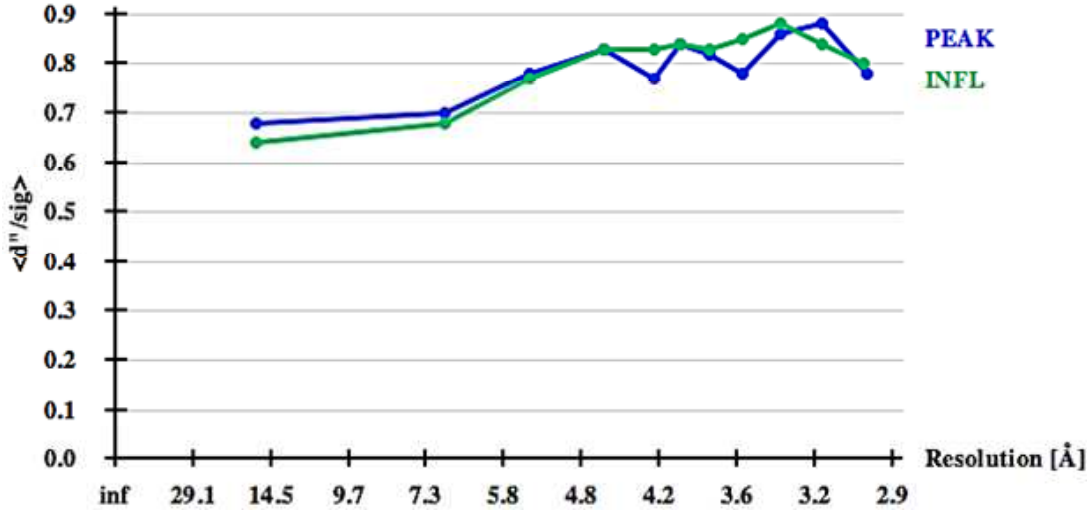
(a) d''/sigma against resolution BPSL1958 K3C datasets



(b) d''/sigma against resolution BPSL1958 S128C datasets



(c) d''/sigma against resolution BPSL1958 A357C datasets



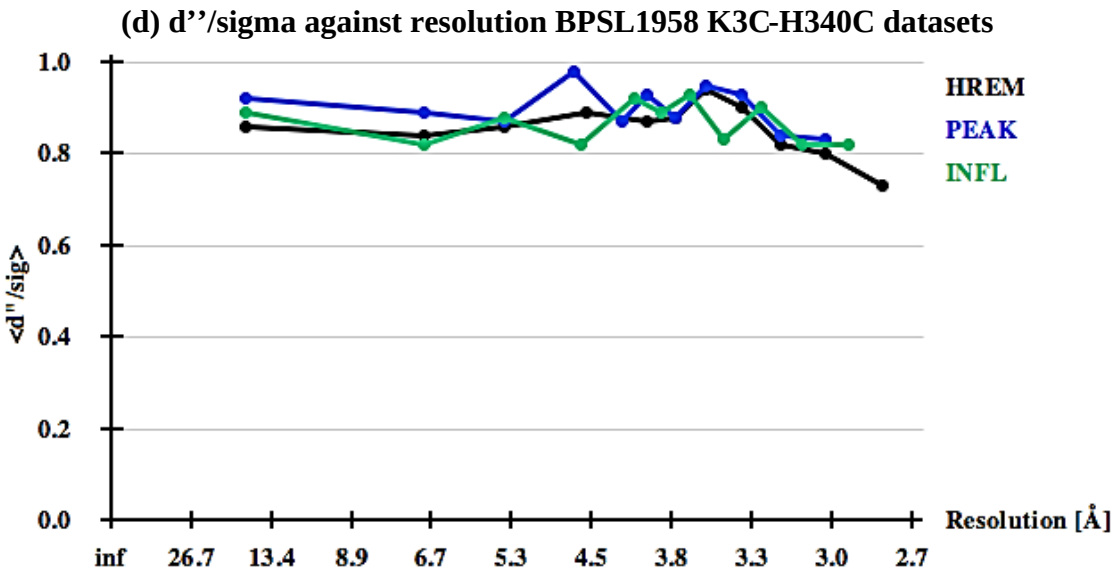


Figure 6.12 Results from SHELX C showing anomalous signal from the four BPSL1958 mutants datasets. The four graphs show d''/sigma plotted against resolution for each dataset. A value of 1.2 or above indicates the presence of good anomalous signal and it can therefore clearly be seen the only datasets that include useful anomalous data are the K3C mutant datasets.

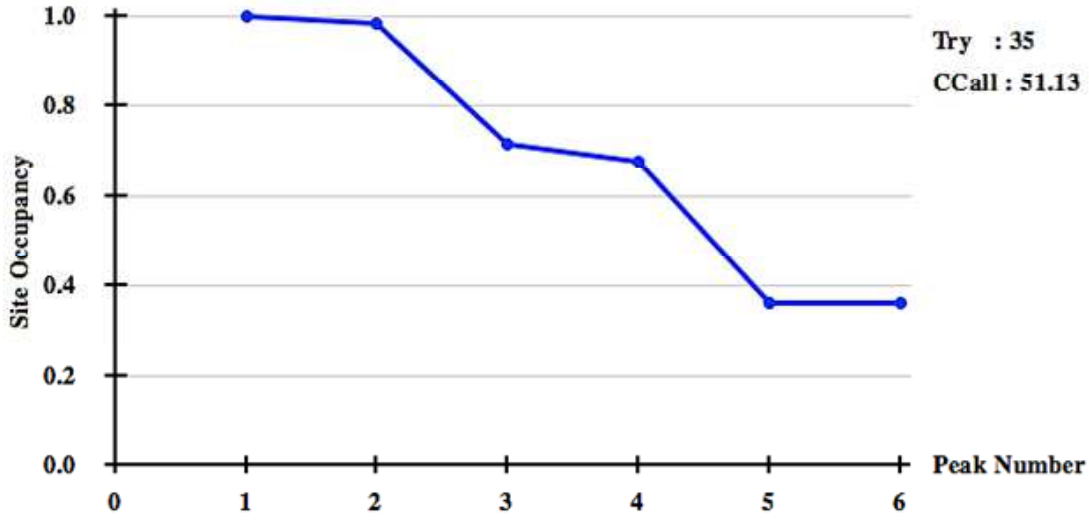
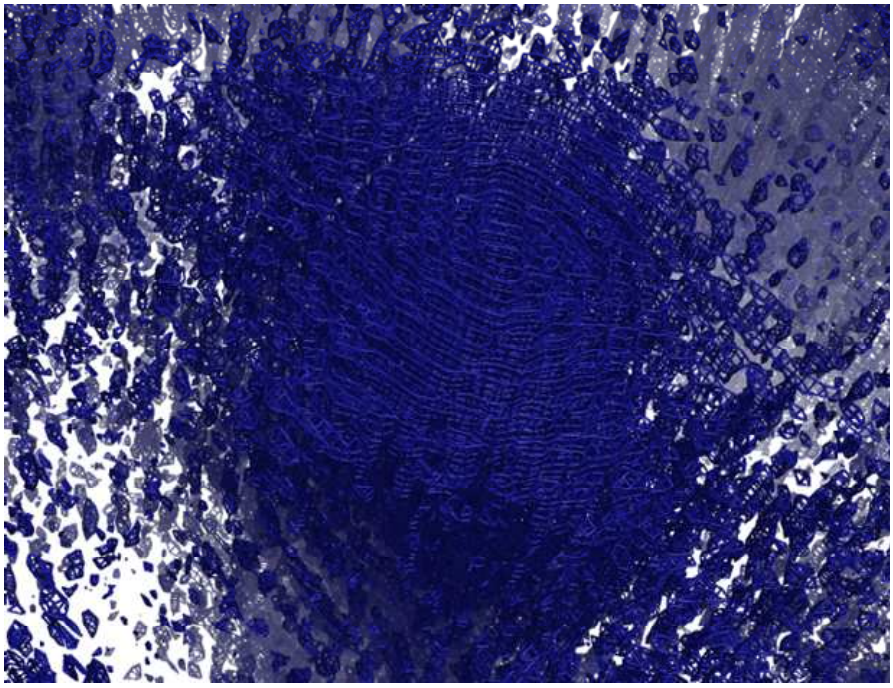


Figure 6.13 Results from SHELX D for BPSL1958 K3C MAD experiment showing the best solution. Peaks are predicted to have occupancies of 0.99, 0.97, 0.71, 0.67, 0.36 and 0.35 and clearly exist as sets of two sites.

(a) Original substructure hand



(b) Inverted substructure hand

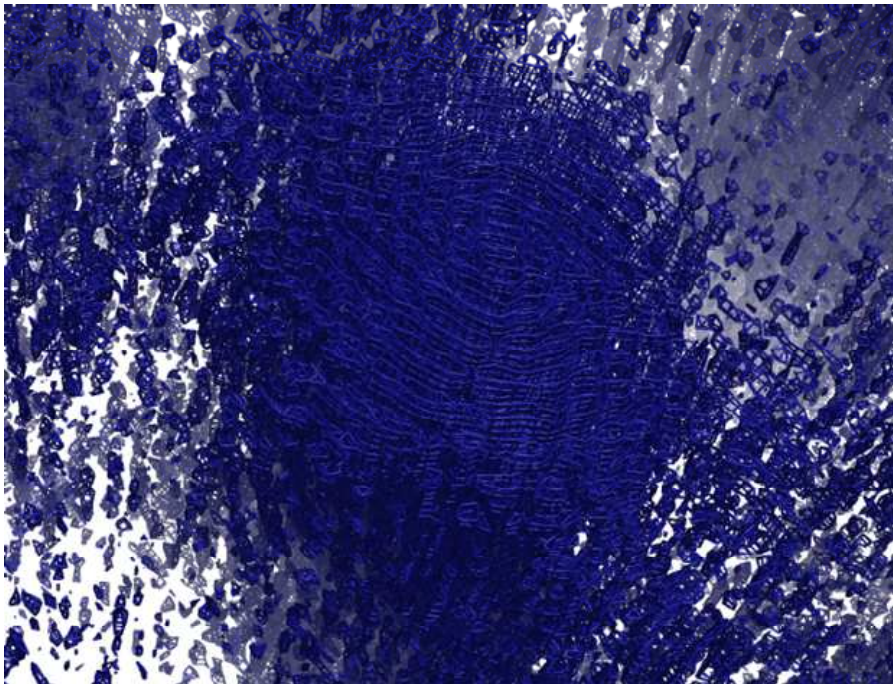


Figure 6.14 Sample region of electron density for the original hand and inverted hand solutions for the BPSL1958 K3C mercury MAD phasing experiment. Both maps are contoured at 1 sigma and show the same cross section of density. **a** The initial electron density map output from SHELX E for the original substructure hand. **b** The initial electron density map output from SHELX E for the inverted substructure hand.

6.7 The space group of BPSL1958 crystals

Analysis of the self-Patterson functions for the native, K3C, S128C and A357C mutant datasets all of which had been assigned to $P2_12_12_1$ suggested they all possessed essentially the same non-crystallographic translational symmetry corresponding to a translation of ½ along the c axis of the unit cell (table 6.9). For the K3C and A357C mutants this was aligned with the other unit cell axes but for the native and S128C mutants it was slightly offset along the a axis. Therefore if the correct space group is $P2_12_12_1$ the asymmetric unit contains two molecules of the protein related to each other by a translation of ½ along the longest unit cell axis.

(a) BPSL1958 native dataset

Peak	Height RMS^{-1}	Fractional coordinates			Orthogonal coordinates		
Origin	135.57	0.0000	0.0000	0.0000	0.00	0.00	0.00
Translation	41.42	0.0220	0.0000	0.5000	1.40	0.00	54.08

(b) BPSL1958 K3C peak dataset

Peak	Height RMS^{-1}	Fractional coordinates			Orthogonal coordinates		
Origin	133.81	0.0000	0.0000	0.0000	0.00	0.00	0.00
Translation	53.67	0.0000	0.0000	0.5000	0.00	0.00	54.58

(c) BPSL1958 S128C peak dataset

Peak	Height RMS^{-1}	Fractional coordinates			Orthogonal coordinates		
Origin	134.81	0.0000	0.0000	0.0000	0.00	0.00	0.00
Translation	41.36	0.0238	0.0000	0.5000	1.52	0.00	54.08

(d) BPSL1958 A357C peak dataset

Peak	Height RMS^{-1}	Fractional coordinates			Orthogonal coordinates		
Origin	106.16	0.0000	0.0000	0.0000	0.00	0.00	0.00
Translation	51.42	0.0000	0.0000	0.5000	0.00	0.00	53.92

Table 6.9 Self Patterson analysis of the native and three single mutant datasets for crystals of BPSL1958. The analyses were conducted using a resolution cut off of 3 Å for each dataset and a grid spacing equivalent to 1 Å along each axis. All datasets suggest the presence of a translation half way along the longest unit cell axis giving rise to peaks equating to 31, 40, 31 and 48 % of the origin peak, for the native, K3C, S128C and A357C datasets respectively.

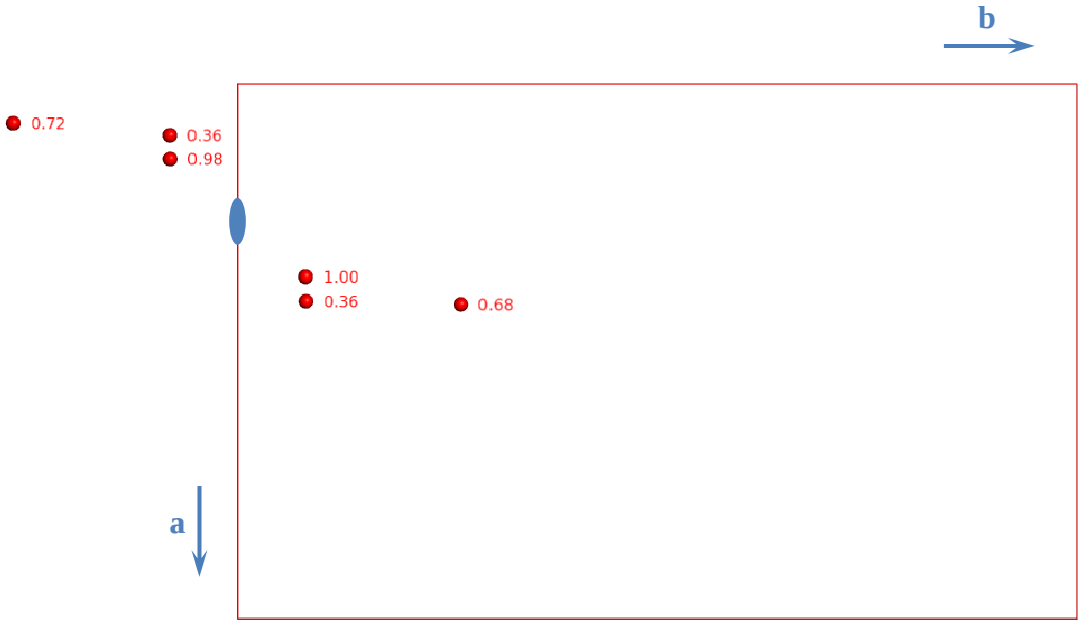


Figure 6.15 Heavy atom sites for the BPSL1958 K3C MAD solution. The image is viewed with the a-axis vertical and the b-axis horizontal to the page clearly showing a non-crystallographic two-fold axis between the sites located at ¼ a and 0 b coincident with the c-axis of the unit cell. The numbers represent the relative occupancies of the sites, and the pairs of sites with similar occupancies are those related by the axis.

Polar angles			**Orthogonal translation**			**Fractional translation**		
Omega	Phi	Kappa	a	b	c	a	b	c
179.95	43.93	179.82	31.48	197.87	0.10	0.4981	2.0001	0.0001

Table 6.10 The non-crystallographic symmetry axis between heavy atom sites for the BPSL1958 K3C MAD solution. The axis was located using Profess conducted with a maximum search distance of 40 Å. As the omega angle is effectively 180°, the rotation around phi is redundant and the resulting rotation is around an axis located coincident with the c-axis of the unit cell. The value of effectively 180° for the kappa angle represents a two-fold rotation around this axis. The translational relationship can be approximated to a translation of ½, 2 and 0 along the axes a, b and c respectively. The a-axis translation could equally be defined as 0, ¼ or ¾ as the space group symmetry operators combine with the NCS axis. The b-axis value of 2 could be more simply defined as 0. This confirms the positioning of the axis of rotation, ¼ along a, and coincident with the b and c unit cell axes found by initial inspection of the heavy atom sites.

On inspection, the heavy atom sites found in the MAD phasing experiment using the K3C mutant datasets appeared to possess a two-fold non-crystallographic symmetry element coincident with the two-fold screw axis along the c-axis of the unit cell (figure 6.15). The programme Profess from the CCP4 suite [156] was used to determine the exact relationship, confirming the presence and location of the non-crystallographic rotation axis (table 6.10). This NCS symmetry element further explains the relationship between two copies of the protein found in the asymmetric unit and the space group of the crystals. As the pseudo symmetry axis is coincident with the crystallographic axis, the asymmetric unit can be considered to consist of two molecules related by a pseudo 2-fold symmetry axis around a crystallographic 2_1 screw axis, or alternatively if the true space group is $P2_12_12$, two molecules related by a translation of ½ along a crystallographic 2-fold axis. The BPSL1958 K3C-H340C double mutant crystallised in a different space group to the others which was monoclinic as opposed to orthorhombic containing a single copy of the molecule.

7.0 Studies on the protein BPSS0211

This section describes the purification, crystallisation and structure solution of the target BPSS0211 by selenium MAD experiments. The section also includes a discussion of the structure of the protein.

7.1 Protein purification for BPSS0211

7.1.1 Native protein purification

Approximately 3 g of cell paste was resuspended in 30 ml 50 mM TRIS pH 8.0 and disrupted by sonication. Cell debris and insoluble proteins were removed by centrifugation at 70,000 g for 15 minutes and the supernatant was loaded onto a DEAE-Sepharose fast flow column equilibrated with 50 mM TRIS pH 8.0. A 200 ml gradient from 0 to 500 mM NaCl was then applied to the column and 8 ml fractions were collected. Fractions were analysed by SDS-PAGE and BPSS0211 was found to not interact with the column coming through in the initial flow-through. The DEAE flow-through was taken and subjected to an ammonium sulphate cut. An equal volume of 4 M ammonium sulphate was added to bring the concentration of ammonium sulphate to 2 M and the solution was incubated on ice for 10 minutes. Precipitated protein was collected by centrifugation at 70,000 g for 10 minutes before resuspension in 20 ml 1.5 M ammonium sulphate, incubation on ice for 10 minutes and centrifugation at 70,000 g for 10 minutes. The supernatant was taken and proteins were precipitated by bringing the overall concentration of ammonium sulphate to 2 M followed by incubation on ice for 10 minutes. The sample was centrifuged at 70,000 g for 10 minutes and the pellet was resuspended in 1 ml 50 mM TRIS pH 8.0 before being loaded on a 1.6 x 60 cm Superose 6 gel filtration column equilibrated in 50 mM TRIS pH 8.0 and 500 mM NaCl and eluted using the same buffer collecting 2 ml fractions (figure 7.1). Peak fractions containing the protein were pooled and concentrated using a Vivaspin concentrator with a 5 kDa MWCO to 8 mg ml^{-1} protein. The buffer was exchanged for 10 mM TRIS pH 8.0 using a Zeba spin desalting column (Thermo scientific) for use in crystallisation trials and the purification was analysed by PAGE (figure 7.2). The overall yield of protein was high with 23 mg being obtained which was estimated by SDS-PAGE to be over 95 % pure.

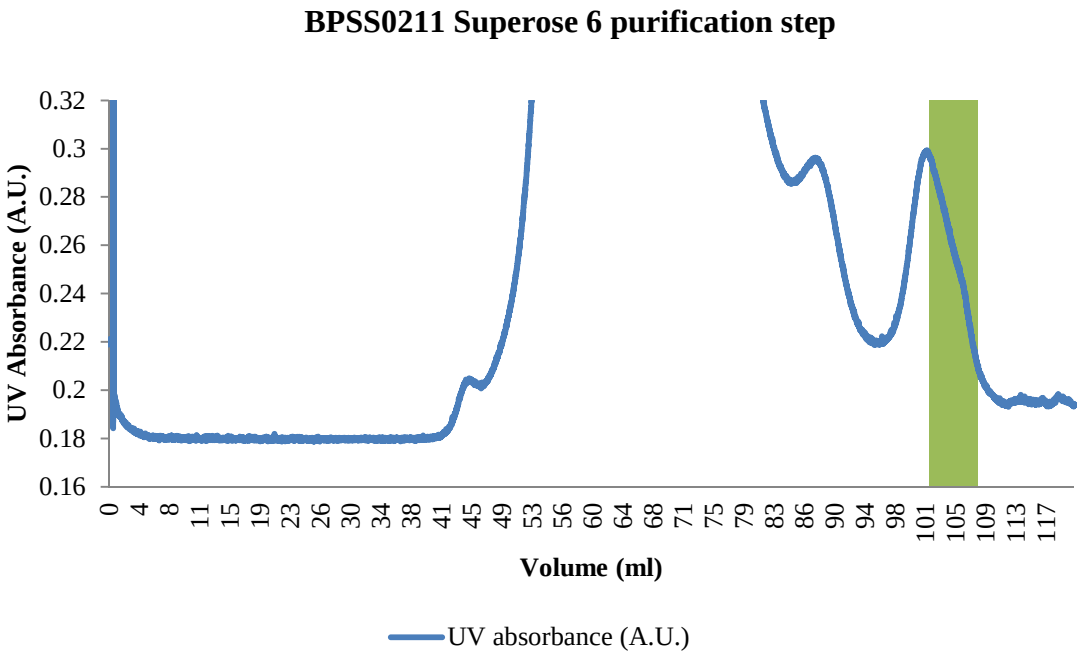


Figure 7.1 Chromatogram trace for the gel filtration purification step of BPSS0211. 2 ml fractions were collected throughout and the green highlighted region indicates fractions taken as pure protein.

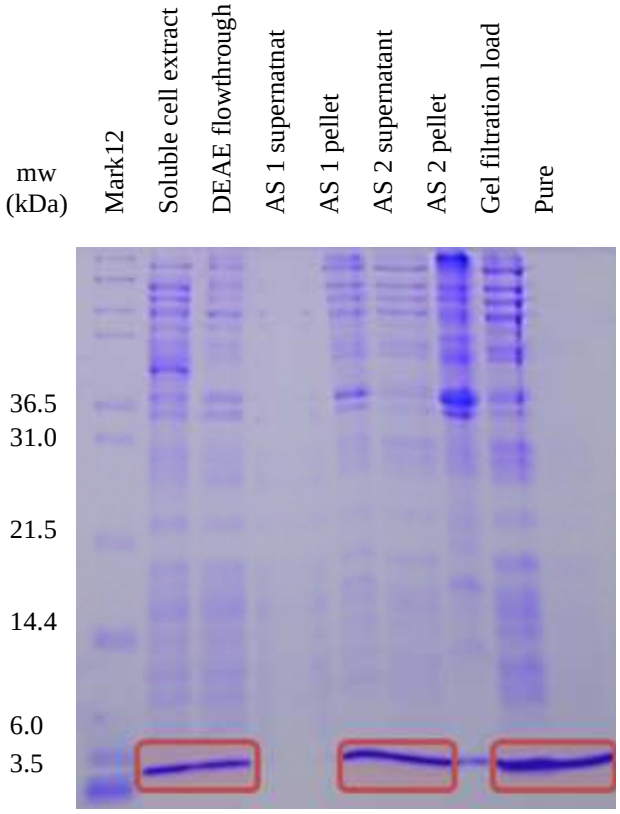


Figure 7.2 SDS-PAGE gel showing the purification of BPSS0211. The molecular weight of BPSS0211 is 6.8 kDa and the highlighted bands indicate the protein in fractions taken for subsequent purification steps or as pure protein.

7.1.2 Seleno-methionine protein purification

Protein containing seleno-L-methionine was purified using the same techniques as for native protein. Approximately 4 g of cell paste was used resulting in a yield of 12 mg protein at 8 mg ml^{-1} for use in crystallisation trials.

7.1.3 Purification analysis

The success of the somewhat simplistic purification protocol arises from the initial high level of overexpression, the protein's inability to interact with DEAE and the protein's relatively small size allowing it to be significantly separated from other proteins by gel filtration. The elution profile for BPSS0211 from the gel filtration column shows a shouldered double peak suggesting differing oligomeric states roughly corresponding to molecular weights of 11 kDa and 23 kDa, which themselves are best explained by a dimeric and tetrameric form of the protein, although this remains to be confirmed (figure 7.1).

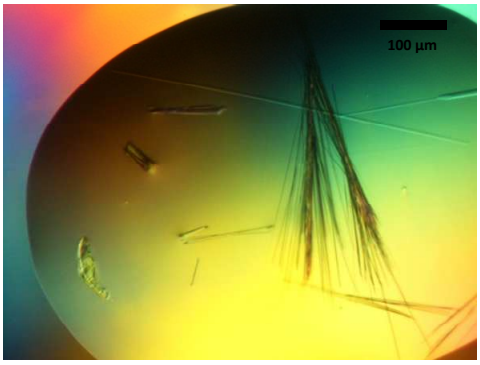
7.2 Protein crystallisation for BPSS0211

7.2.1 Native protein crystallisation

Four initial 96 condition robot screens, the JCSG+, PACT, Pegs and Classics suites, were conducted using purified native protein. 200 nl of protein was mixed with 200 nl of well solution and the trays were incubated at 17 ºC. Several initial hits were obtained, producing a variety of crystal forms across different conditions (figure 7.3), PACT B2 – B5 containing 100 mM MIB buffer varying from pH 5 to 8 over the four conditions and 25 % (w/v) PEG 1500 produced a long rod crystal form, JCSG B6 containing 100 mM phosphate-citrate buffer pH 4.2, 40 % (v/v) ethanol and 5 % (w/v) PEG 1000 produced a cubic form and JCSG E6 containing 100 mM imidazole pH 8.0, 200 mM zinc acetate and 20 % (w/v) PEG 3000 produced a multifaceted crystal form. These conditions were selected for hanging drop optimisation trials, PACT B2 – B5 were optimised by varying the PEG concentration (15 – 30 % (w/v)) and pH (5.0 – 8.0), JCSG B6 by altering PEG concentration (0 – 20 % (w/v)), ethanol concentration (20 – 50 % (v/v)) and pH (4 – 6) and JCSG E6 by varying PEG concentration (5 – 30 % (w/v)) and pH (7.5 – 8.5). The optimisations for PACT B2 – B5 and JCSG B6 did not produce any larger crystals of better quality than the robot screen while attempts to optimise around the JCSG E6 condition led to the production of large, well defined crystals.

7.2.2 Seleno-methionine protein crystallisation

Seleno-methionine crystals were produced by screening around the optimised JCSG E6 condition varying PEG concentration (5 – 15 % (w/v)) and were found to be best in the same condition as for native.

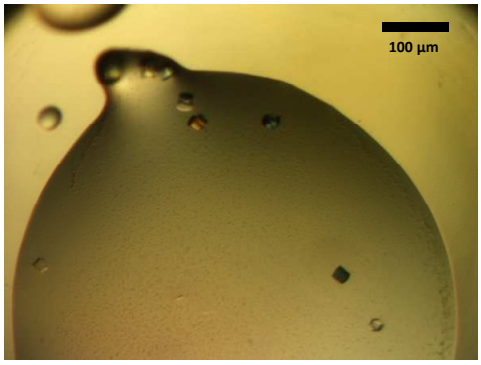


PACT B3 – Robot screen

100 mM MIB pH 6

25 % (w/v) PEG 1500

Drop size 200 nl protein + 200 nl well solution



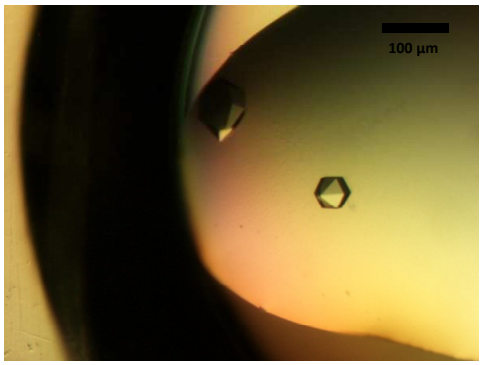
JCSG B6 – Robot screen

100 mM Phosphate-citrate pH 4.2

40 % (v/v) Ethanol

5 % (w/v) PEG 1000

Drop size 200 nl protein + 200 nl well solution



JCSG E6 – Robot screen

100 mM imidazole pH 8

200 mM zinc acetate

20 % (w/v) PEG 3000

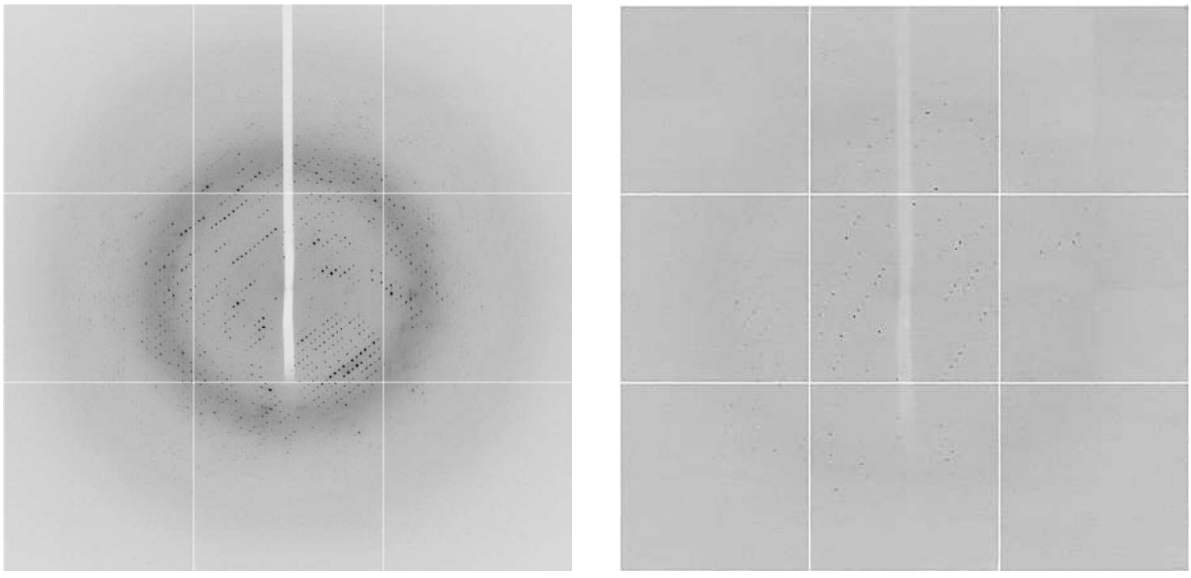
Drop size 500 nl protein + 200 nl well solution

Figure 7.3 Photographs of BPSS0211 native crystals. The crystals in the three different conditions exhibit different crystal forms.

7.3 BPSS0211 Native data

7.3.1 Native data collection

Native crystals were selected for data collection based on their size and overall definition from the optimisation trials. These were from the optimised JCSG E6 condition, 100 mM imidazole pH 8.0, 200 mM zinc acetate and 12 % (w/v) PEG 3,000. Crystals were looped and placed into a cryoprotectant solution consisting of the well solution with the addition of 30 % (v/v) ethylene glycol. They were then looped again and flash frozen in a stream of gaseous nitrogen at 100 ºK and mounted onto the detector. Initial diffraction analysis was conducted in order to determine the diffraction quality of several crystals using an in-house Rigaku MM007 rotating anode generator and a MAR 345 research image plate. Two images were collected 90º apart with 1º oscillation. A number of crystals that diffracted beyond 3 Å on the home source were saved and taken to a synchrotron at the I03 beamline of the Diamond light source, Oxford. Two initial test images were taken 90º apart with 1º oscillation to ensure crystal centering and to obtain a collection strategy using the auto-indexing and collection strategy components of Mosflm [146]. For the data set 100 images were collected with 1º oscillation per image using X-rays of 12678 eV at a crystal to detector distance of 256 mm using an ADSC Q315r detector. Data extending to 2.2 Å were collected (figure 7.4).



Native dataset **Seleno-methionine peak dataset**

Figure 7.4 Diffraction images of native and seleno-methionine crystals of BPSS0211.

7.3.2 Native data processing

The datasets were auto-processed at Diamond through the xia2 system using the 3dii mode [147]. Data were indexed and integrated by XDS and scaled by XSCALE [148]. Data indexed to space group $P6_222$ or $P6_422$, with the unit cell parameters a = b = 69.3 Å, c = 84.1 Å, (table 7.1). Matthews coefficients calculated using Mattprob [149] showed the asymmetric unit was predicted to contain one, two or three protein molecules giving Matthews coefficients of 4.29, 2.15 or 1.43 respectively, with two molecules being the most probable (table 7.2).

	Native protein crystal	**Seleno-methionine protein crystal**		
		peak	**inflection**	**high energy remote**
Space group	P 6_2 2 2	P 6_2 2 2	P 6_2 2 2	P 6_2 2 2
Unit cell parameters				
a (Å)	69.3	69.2	69.3	69.4
b (Å)	39.3	69.2	69.3	69.4
c (Å)	84.1	83.9	84.1	84.2
Energy (eV)	12678	12661	12655	13450
Resolution range (Å)	59.99 – 2.17	59.91 – 3.02	60.04 – 3.08	60.07 – 3.13
Unique reflections	6751 (480)	2605 (178)	2485 (182)	2380 (166)
R_{merge}	0.073 (0.682)	0.147 (1.072)	0.151 (0.814)	0.147 (0.723)
R_{pim}	0.025 (0.215)	0.045 (0.303)	0.046 (0.235)	0.044 (0.207)
Completeness (%)	99.7 (99.5)	99.7 (99.4)	99.8 (100.0)	99.7 (100.0)
Anomalous completeness (%)	99.7 (99.0)	99.7 (99.3)	99.8 (100.0)	99.8 (100.0)
Multiplicity	11.1 (11.7)	12.9 (14.1)	12.9 (13.5)	12.9 (13.8)
Anomalous multiplicity	6.2 (6.3)	7.5 (7.6)	7.5 (7.7)	7.5 (7.6)
Mean (I)/σ(I)	18.9 (3.8)	16.3 (3.1)	15.5 (4.0)	16.7 (4.3)

Table 7.1 Data collection statistics for native and seleno-methionine BPSS0211 crystals. Numbers in parentheses indicate values for the highest resolution shell.

Molecules in the AU	**Probability (based on resolution)**	**Probability (based on all proteins)**	**Vm ($Å^3$ / Da)**	**Solvent content (%)**	**Molecular weight (Da)**
1	0.0033	0.0572	4.29	71.33	6794
2	0.9936	0.9367	2.15	42.66	13588
3	0.0023	0.0060	1.43	13.99	20382

Table 7.2 Matthews coefficient calculations and probabilities for BPSS0211. The results show a possibility of one, two or three protein molecules inhabiting the asymmetric unit.

7.4 BPSS0211 Seleno-methionine protein data

7.4.1 Seleno-methionine data collection

Seleno-methionine crystals were selected from the same condition as for native crystals and looped and mounted using the same methodology. A number of crystals were saved and data was collected from a single crystal using a synchrotron at the I03 beamline of the Diamond light source, Oxford. A fluorescence scan at the selenium K edge was conducted to confirm Selenium incorporation and select X-ray energies at which to collect data in a MAD experiment (figure 7.5 a). Three energies were selected for data collection based on the fluorescence absorbance spectrum. The peak (maximised f'') and inflection point (minimised f') were selected as 12661 eV and 12655 eV respectively using CHOOCH [157] (figure 7.5 b) and the high energy remote (maximised Δf' from inflection) was selected as 13450 eV to provide a similar beam size and flux. Two initial test images were taken 90º apart with 1º oscillation to ensure crystal centering and to obtain a collection strategy using the auto-indexing and collection strategy components of Mosflm [146]. For each wavelength 120 images were collected with 1º oscillation per image at a crystal to detector distance of 375.3 mm using an ADSC Q315r detector. Data extending to 3.1 Å were collected for each X-ray energy (figure 7.4).

7.4.2 Seleno-methionine data processing

The X-ray diffraction datasets were auto-processed at Diamond through the xia2 system using the 3dii mode [147]. Data were indexed and integrated by XDS and scaled by XSCALE [148]. Data indexed to the same space group as for native data with similar unit cell parameters (table 7.1).

7.5 Experimental phasing for BPSS0211

Initial phasing was conducted with the programmes of the SHELX package [154] using the HKL2MAP graphical user interface for SHELXC and SHELXD [155]. The native protein crystal dataset along with the seleno-methionine protein crystal peak, inflection and high energy remote datasets were inputted into SHELXC. The seleno-methionine datasets contained anomalous signal beyond 4.5 Å, and surprisingly the native dataset was also found to contain significant anomalous signal to beyond 3.0 Å (figure 7.6). The anomalous signal in the native data was found to arise from the incorporation of metal ions in the crystal structure. SHELXD was used to calculate the heavy atom substructure using all data to 3.5 Å. The best

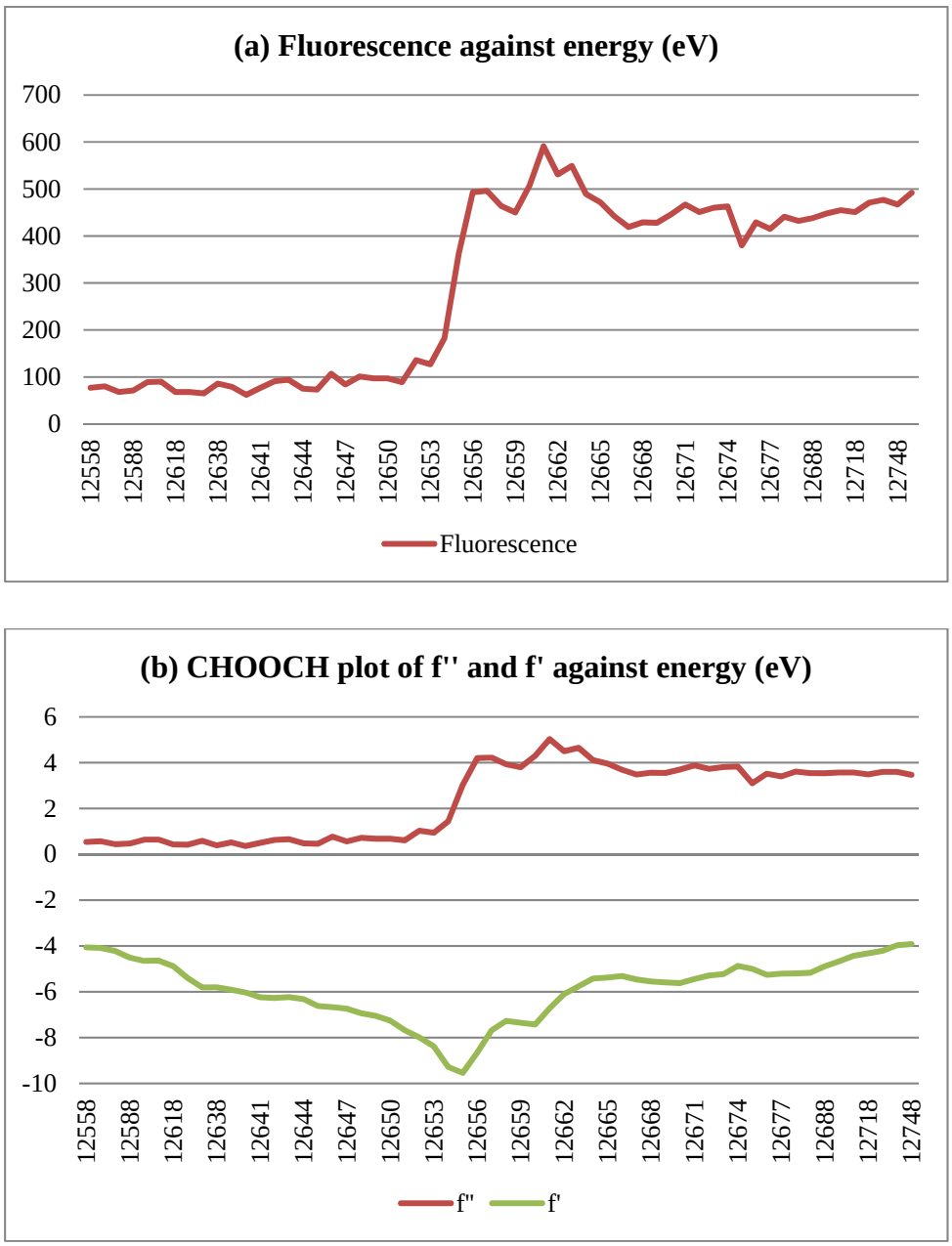
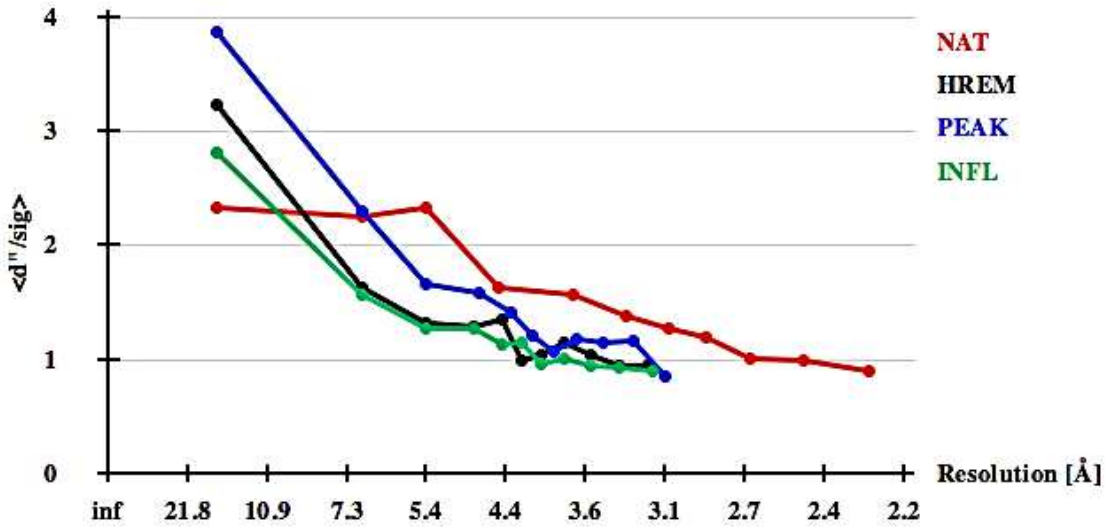


Figure 7.5 Selenium K-edge fluorescence scan and CHOOCH plot for BPSS0211 seleno-methionine crystals. **a** The fluorescence spectrum at the Selenium K-edge confirmed the incorporation of Selenium into the protein crystals and showed maximal fluorescence at 12661 eV, 3 eV above the theoretical value of 12658 eV. **b** Three energies were chosen for the MAD experiment based on the CHOOCH plot, the peak to obtain maximum f'', the inflection point to obtain minimum f', and the high energy remote to maximise Δf'. The energies used were 12661 eV for the peak, 12655 eV for the inflection and 13450 eV for the high energy remote.

solution was identified with a correlation coefficient of 44.65 and a Patterson figure of merit of 16.60 in space group $P6_222$. A total of three potential heavy atom sites were found with one on the 2-fold axis and therefore lying on a special position (figure 7.7). Preliminary protein phasing, density modification and initial model building were conducted using

(a) d''/sigma against resolution



(b) Self-anomalous correlation coefficient against resolution

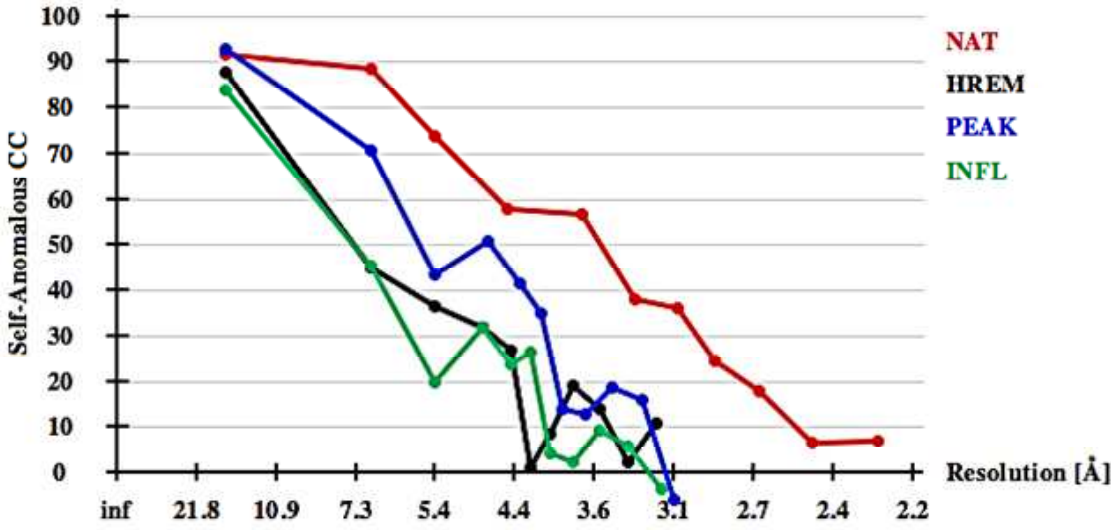


Figure 7.6 Results from SHELX C showing anomalous signal from the four BPSS0211 datasets. **a** Graph of d''/sigma plotted against resolution. A value of 1.2 or above indicates of good anomalous signal. **b** Graph of self-anomalous correlation coefficient plotted against resolution. A value of 30 % or above indicates good anomalous signal.

SHELXE-beta with auto-tracing. Five rounds of twenty cycles of phase calculation and density modification followed by auto-tracing were carried out for both the inverted and the original hands of the selenium substructure. A solvent content of 71, 43 and 14 % were used corresponding to one, two or three molecules in the asymmetric unit. The best solution gave

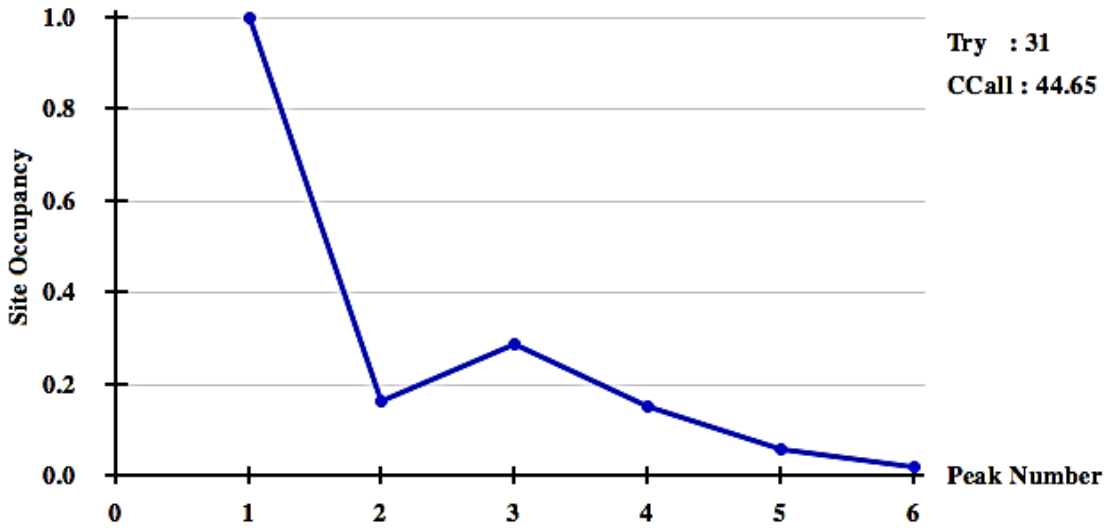


Figure 7.7 Results from SHELX D for BPSS0211 MAD experiment showing the best solution. Peaks are predicted to have occupancies of 1.00, 0.32 (on a special position so halved to 0.16) and 0.28.

values of 1.220 and 0.831 for the contrast and connectivity and a model consisting of 51 residues in two alpha helices was built with a correlation coefficient for partial structure against native data of 48.87 %. These results (table 7.3) coupled with an inspection of the resultant electron density maps (figure 7.8 a) proved the original hand with a solvent content of 71 % was the correct solution.

Hand		**Original**			**Inverted**	
Predicted molecules in the ASU	1	2	3	1	2	3
Solvent content	0.71	0.43	0.14	0.71	0.43	0.14
Contrast	1.220	1.074	1.684	0.609	0.419	0.201
Connectivity	0.831	0.803	0.736	0.737	0.700	0.658
Pseudo-free correlation coefficient	76.14	73.99	58.15	45.15	51.54	47.45
Residues built	51	53	51	23	-	-
Correlation coefficient for structure	48.87	48.92	46.98	8.97	-	-
Mean figure of merit for structure	0.730	0.717	0.548	0.43	-	-

Table 7.3 Results from SHELX E for BPSS0211 MAD experiment. The programme was run for different solvent contents based on the presence of one, two or three protein molecules in the asymmetric unit for the original and the inverted hand. SHELX E failed to trace anything for the inverted hand solutions with two or three molecules in the asymmetric unit. The best solution is for one molecule in the asymmetric unit with the original hand giving a significantly better result.

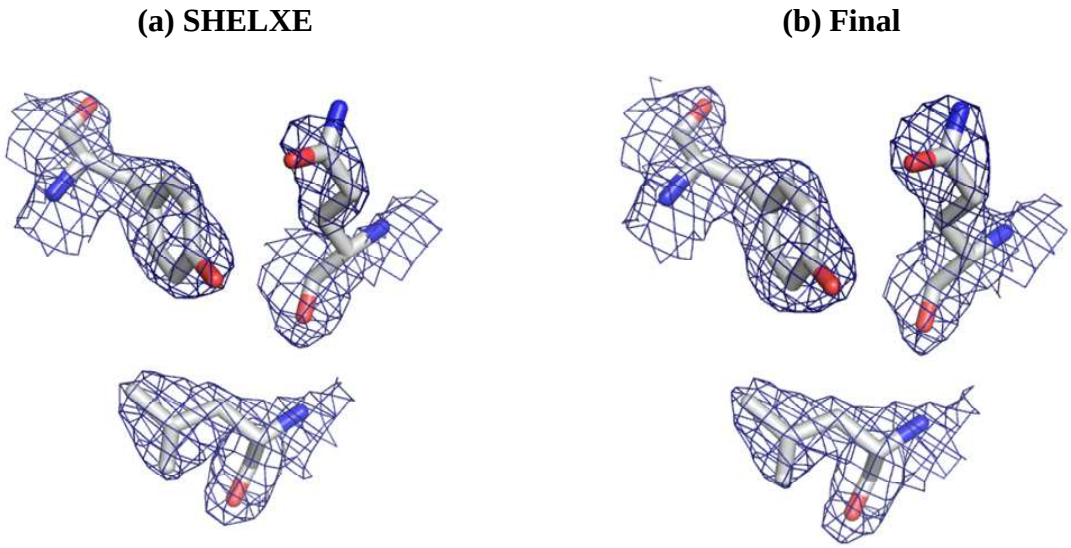


Figure 7.8 Sample region of electron density for BPSS0211. Both maps use data to 2.2 Å, are contoured at 1.5 sigma and show the same region of density around residues Tyr-12, Gln-34 and Leu-38 from the final model. **a** The initial electron density map output from SHELX E. **b** The electron density map for the final model.

7.6 Model Building and refinement for BPSS0211

A structure containing side chains was constructed using AutoBuild, part of the PHENIX suite of programs [158]. The initial poly alanine chain produced by SHELXE was used as a starting model along with the native data and the primary sequence of the protein. The resulting model consisted of 51 residues, with side chains, arranged in 2 alpha helical fragments with 32 water molecules. The model had an R-factor of 0.2949 (R_{free} 0.3126) and a correlation coefficient of 0.68. Further refinement was conducted using iterative cycles of model building, refinement and evaluation using Coot [159] and REFMAC5 [152]. For the final model all water molecules were removed and a select few were included using strict criteria to avoid over fitting the density. Water molecules were only placed in positions where there was a sigma level of 2.0 or above giving a molecule with a b-factor of less than 60 $Å^2$ and sensibly positioned in relation to hydrogen bonding to the protein or other ligands. The final model consists of 51 residues in a single chain, 2 metal ions, 1 imidazole, 1 acetate ion and 26 water molecules (figure 7.9). The model has an R-factor of 0.2485 (R_{free} 0.2996) and agrees well with the electron density (figure 7.8 b).

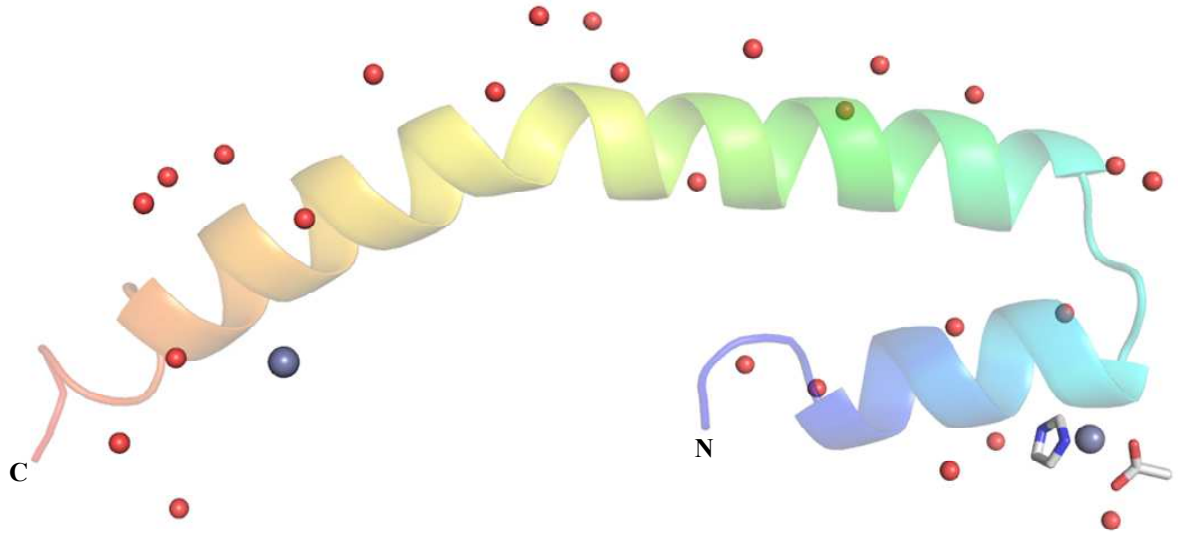
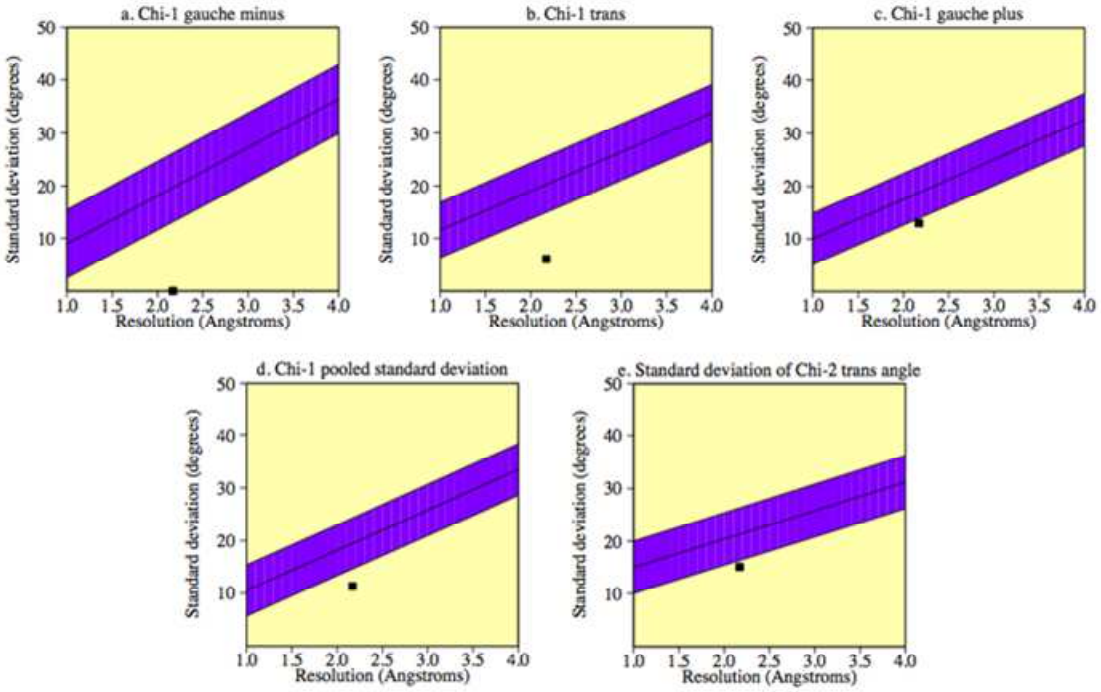


Figure 7.9 Cartoon representation of the overall fold for the final structure of BPSS0211. The protein is coloured as a chainbow from blue to red, the imidazole and acetate molecules are shown as sticks , red spheres represent water molecules and the grey spheres represent metal ions.

Resolution (Å)	2.17
Number of reflections	6408
Protein molecules per asymmetric unit	1
Number of atoms	422
Number of residues	51
Number of waters	26
Number of ions	2
Ramachandran favoured (%)	95.9
Ramachandran outliers (%)	0.0
RMS bond length deviation (Å)	0.014
RMS bond angle deviation (º)	1.572
Average main chain B-factors ($Å^2$)	36.9
Average side chain B-factors ($Å^2$)	48.5
Average waters B-factors ($Å^2$)	49.0
Average buffer component B-factors ($Å^2$)	40.2
R-factor (%)	0.249
R_{free} (%)	0.299
Molprobity score	1.45 (99^{th} percentile)

Table 7.4 Final refinement and validation statistics for BPSS0211.

(a) Side chain parameters



(b) Main chain parameters

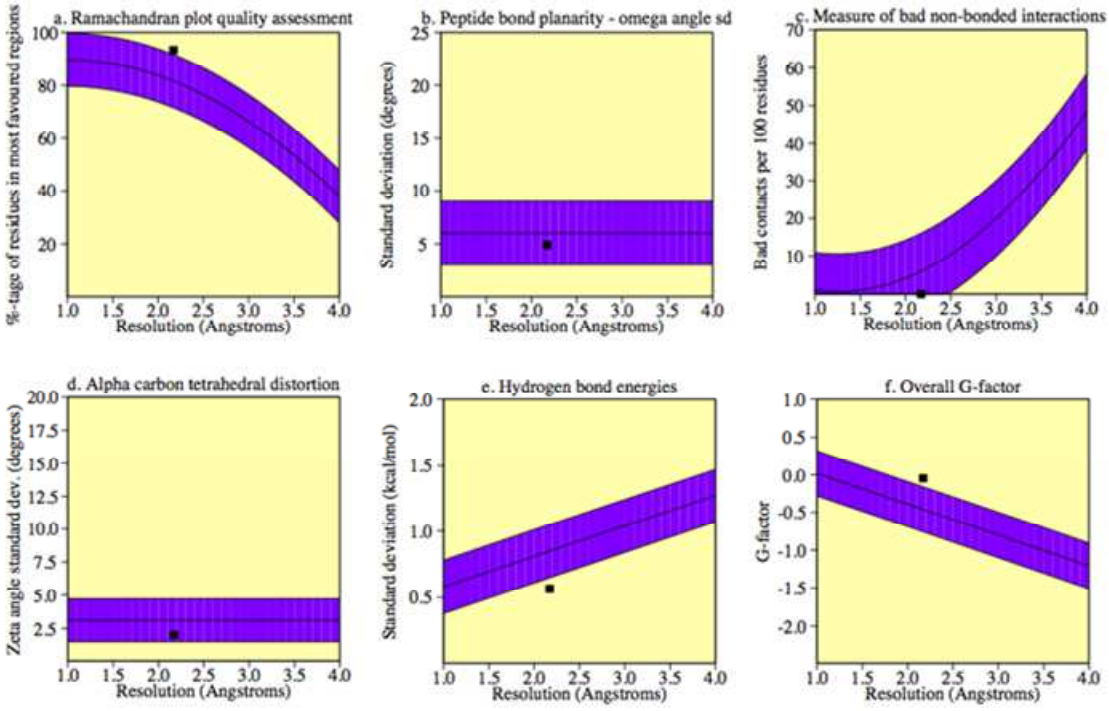


Figure 7.10 Main chain and side chain properties for the final BPSS0211 model. The figure was generated using PROCHECK [157] and shows all residues have properties better than or within the expected range for the resolution of the data.

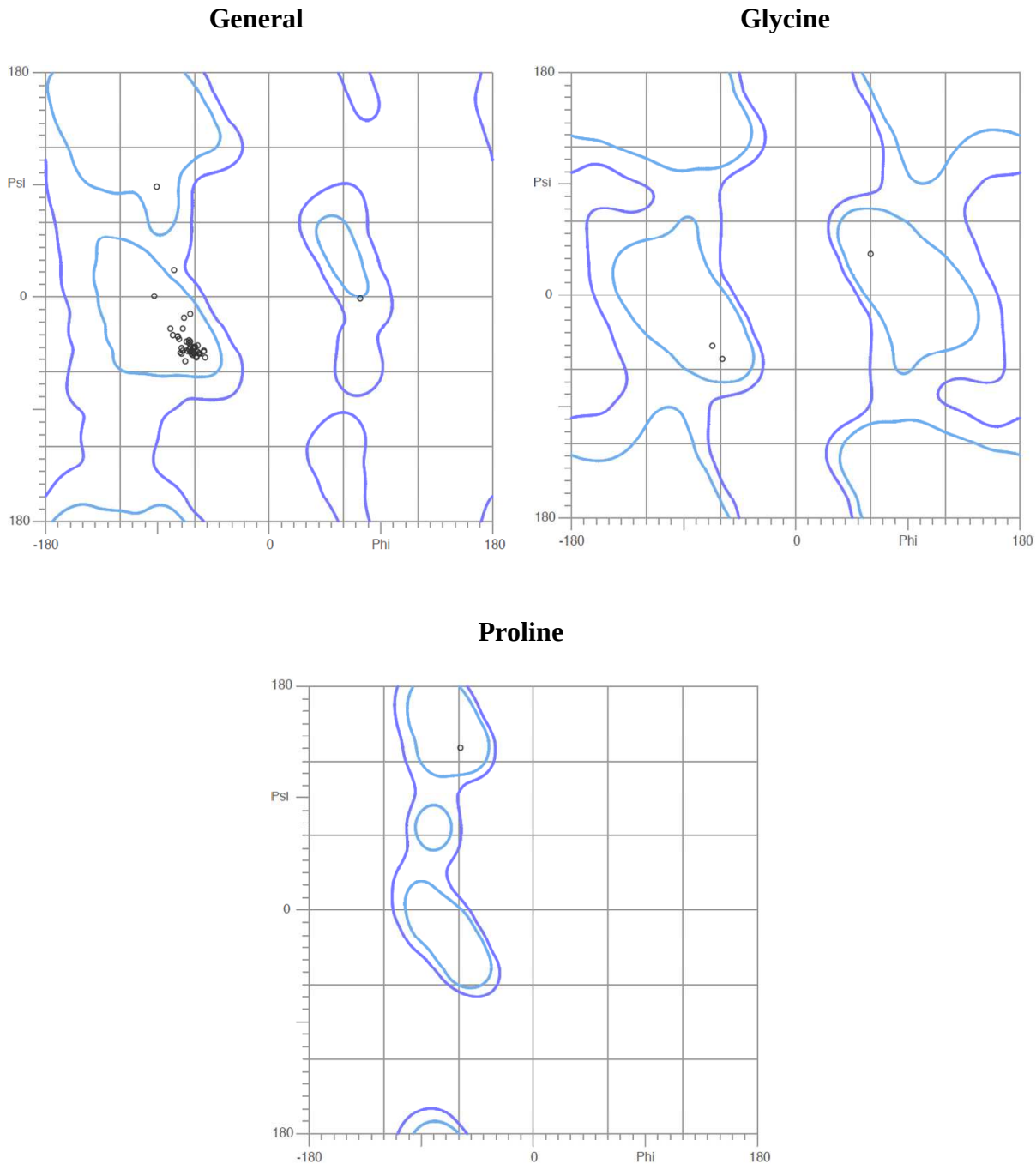


Figure 7.11 Ramachandran plot and statistics for the final BPSS0211 model. The figure was generated using Molprobity [158] and shows all residues have acceptable values for their phi and psi angles with 95.9 % falling within favoured regions.

7.7 The model of BPSS0211

The model has been restrained to standard bond lengths and angles, and the root mean square deviation (RMSD) of bond lengths and angles in the final structure is 0.014 Å and 1.572º, respectively. The structure was validated using PROCHECK [160] and the Molprobity server [161] which showed the overall structure was of good quality (table 7.4). All main chain and side chain parameters were better than or within the expected range for the resolution of the data (figure 7.10) and all residues fell within allowed regions of the Ramachandran plot (figure 7.11).

7.7.1 Alternate residue conformation in the BPSS0211 structure

A single residue in the structure, Glu-55, has been modelled as having alternative conformations with 50 % occupancy in both forms as this best fits the electron density (figure 7.12).

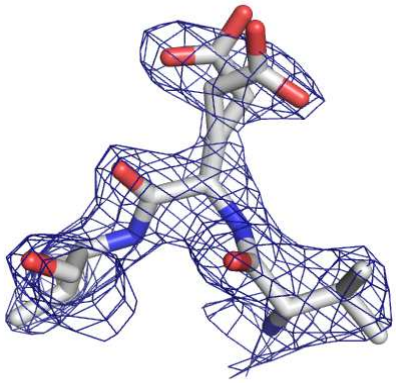


Figure 7.12 The alternate conformations of Glu-55 in the BPSS0211 structure. The map is contoured to 1.5 sigma and despite the residue not being entirely contained within the electron density the dual conformation best explains the density.

7.7.2 Metal ions in the BPSS0211 structure

The final model contains two metal ions, which due to the presence of 200 mM Zn^{2+} in the crystallisation solution were assumed to be zinc although no analysis was done to confirm this assumption. One of the zinc ions (Zn 2 site) is thought to be potentially biologically significant as it forms interactions present on a dimer interface and is co-ordinated by Glu-53 and a symmetry related mate (figure 7.13 a) (section 7.9.2), with the other (Zn 1 site) forming a crystal contact, co-ordinated by His-17, Asp-22 and a single acetate and imidazole molecule from the buffer components (figure 7.13 b).

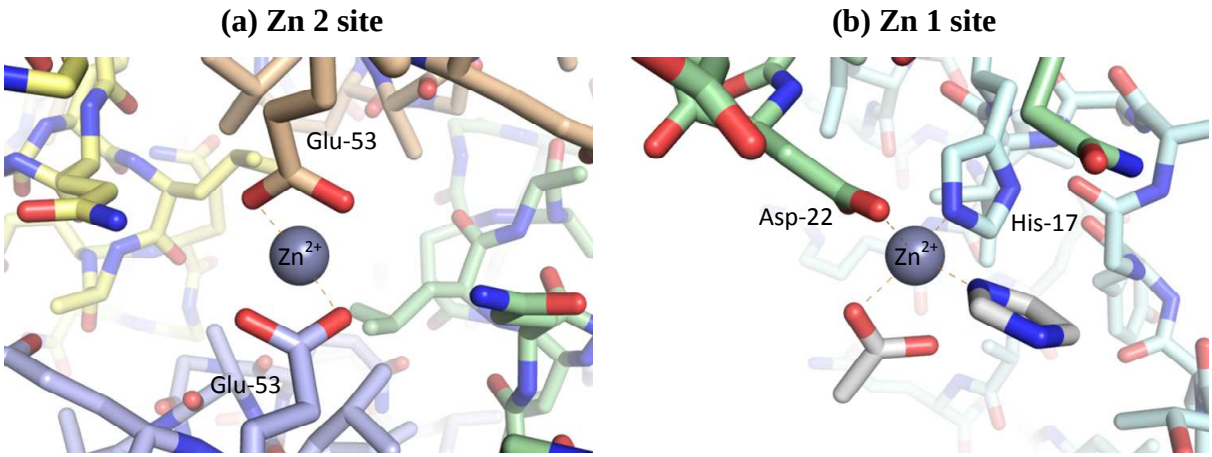


Figure 7.13 Metal ions and their co-ordinating ligands in the BPSS0211 structure. The regions shown include the metal ions (grey spheres), buffer components (white carbon atom structures) and a number of residues from symmetry related monomers of BPSS0211 (represented by the blue, pink, yellow, green and cyan carbon atoms in structures). **a** The Zn 2 site lies on a 2-fold axis and is co-ordinated by Glu-53 residues provided from two symmetry related molecules. **b** The Zn 1 site is co-ordinated by residues His-17 and Asp-22 from two symmetry related models, along with an acetate and imidazole molecule from the buffer components.

7.7.3 Unmodelled density in BPSS0211 structure

The first and last residues for which there is unambiguous density are Thr-10 and Glu-60, respectively. For both termini there are regions of unmodelled density (figure 7.14 a and b) possibly representing further residues unable to be built due to disorder in the crystal.

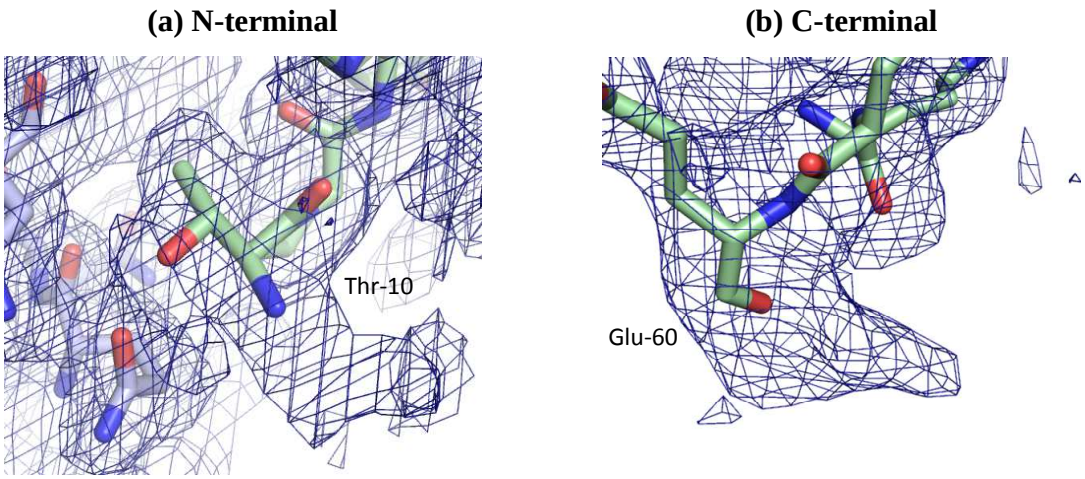


Figure 7.14 Electron density maps showing regions of unmodelled density for the BPSS0211 termini. Both maps are contoured to 0.5 sigma. **a** Unmodelled density at the N-terminal showing the possibility other residues may be present but disordered. **b** Unmodelled density at the C-terminal showing the possibility other residues may be present but disordered.

The final structure contains a number of other regions where interpretation of the density is ambiguous. One of these lies close to a Zn^{2+} site metal ion (figure 7.15). Attempts to explain the broad electron density feature, which had a peak height in the difference map equivalent to a well ordered water molecule, with either water molecules or buffer components were unsatisfactory. Disregarding geometry, the density could be explained with the addition of three water molecules, with one lying on the 2-fold symmetry axis having half occupancy. For data of the resolution and quality obtained it is impossible to interpret this area unambiguously but the most likely explanation is that it represents a mixed population of water molecules and buffer components.

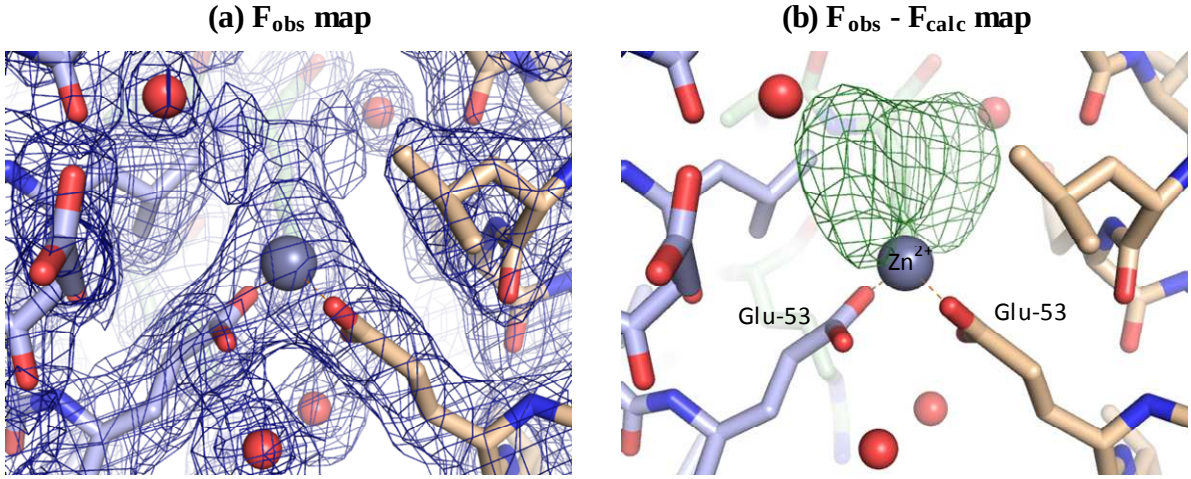


Figure 7.15 Unmodelled density around the zinc 2 metal ion in the BPSS0211 structure. The region shown includes the metal ion (grey sphere), two water molecules (red spheres) and a number of residues from two symmetry related monomers of BPSS0211 (represented by blue and pink carbon atoms in two structures). **a** The F_{obs} map for the density contoured at 0.5 sigma. The extra density above the metal ion clearly represents additional unmodelled molecules co-ordinating the remaining sites of the metal ion. **b** The F_{obs} - F_{calc} map for the density contoured at 5 sigma. The presence of a peak at this level overwhelmingly demonstrates the need for additional components in the model.

7.8 BPSS0211 represents a novel structure

The structure of BPSS0211 was submitted to the Dali server [162] for comparison to known structures in the PDB (table 7.5). The resulting hits had poor Z-scores and levels of sequence identity, and alignments to the known structures were unconvincing. The relatively low RMSD scores observed for the top hits can be explained by the relatively small size of the structure (51 residues) and the helical nature of the protein, limiting the available possible conformations the structure could adopt to those allowed for helix packing and therefore observed in other proteins, negating this as a means of assigning functional significance.

Number	PDB-Chain	Z-score	R.M.S.D	Aligned residues	Length of PDB model	Sequence identity (%)	Molecule description (species)
1	3oaa-O	5.2	2.3	51	284	10	ATP Synthase alpha subunit (*Escherichia coli*)
2	3oaa-W	5.2	2.3	51	284	10	ATP Synthase alpha subunit (*Escherichia coli*)
3	2wpd-G	5.1	2.2	51	269	6	Mitochondrial ATP Synthase alpha subunit (*Saccharomyces cerevisiae*)
4	1vf6-B	5.1	3.2	51	60	12	PALS1-Associated tight junction protein (*Homo sapiens*)
5	4b2q-G	5.1	2.2	51	269	6	Mitochondrial ATP Synthase alpha subunit (*Saccharomyces cerevisiae*)
6	4b2q-g	5.1	2.2	51	269	6	Mitochondrial ATP Synthase alpha subunit (*Saccharomyces cerevisiae*)
7	1kmi-Z	5.0	1.8	51	177	20	CHEY – Chemotaxis protein (*Escherichia coli*)
8	4adz-A	5.0	3.6	51	90	4	CSOR – Copper sensitive operon repressor (*Sterptomyces lividans*)
9	2w6g-G	5.0	2.5	48	140	2	Mitochondrial heart isoform ATP Synthase alpha subunit (*Bos taurus)*
10	4adz-B	5.0	3.7	51	90	4	CSOR – Copper sensitive operon repressor (*Sterptomyces lividans*)

Table 7.5 Dali server results for the model of BPSS0211. The top ten hits are listed alongside their related Z-scores, RMSD scores, alignment statistics and a brief description of each protein. The combination of relatively low Z-scores and poor alignments suggest that BPSS0211 is not homologous to any structure currently in the PDB.

7.9 Analysis of the quaternary structure of BPSS0211

Analysis of the crystal packing shows that the subunits form plausible dimers and tetramers. If these are biologically relevant they might confirm the apparent oligomeric states predicted from the gel filtration profile (figure 7.1). Based on the characteristics of the different monomer-monomer interfaces the quaternary structure can best be described as a dimer of dimers. Subunits A (green) and B (blue) form one dimer and the other two subunits C (yellow) and D (pink) form the other dimer which in turn interact to form the tetramer (figure 7.16). The assemblies and interfaces between monomers were explored using the PISA webserver [163] with a model containing only the protein chain and the zinc ion present on the 2-fold symmetry axis.

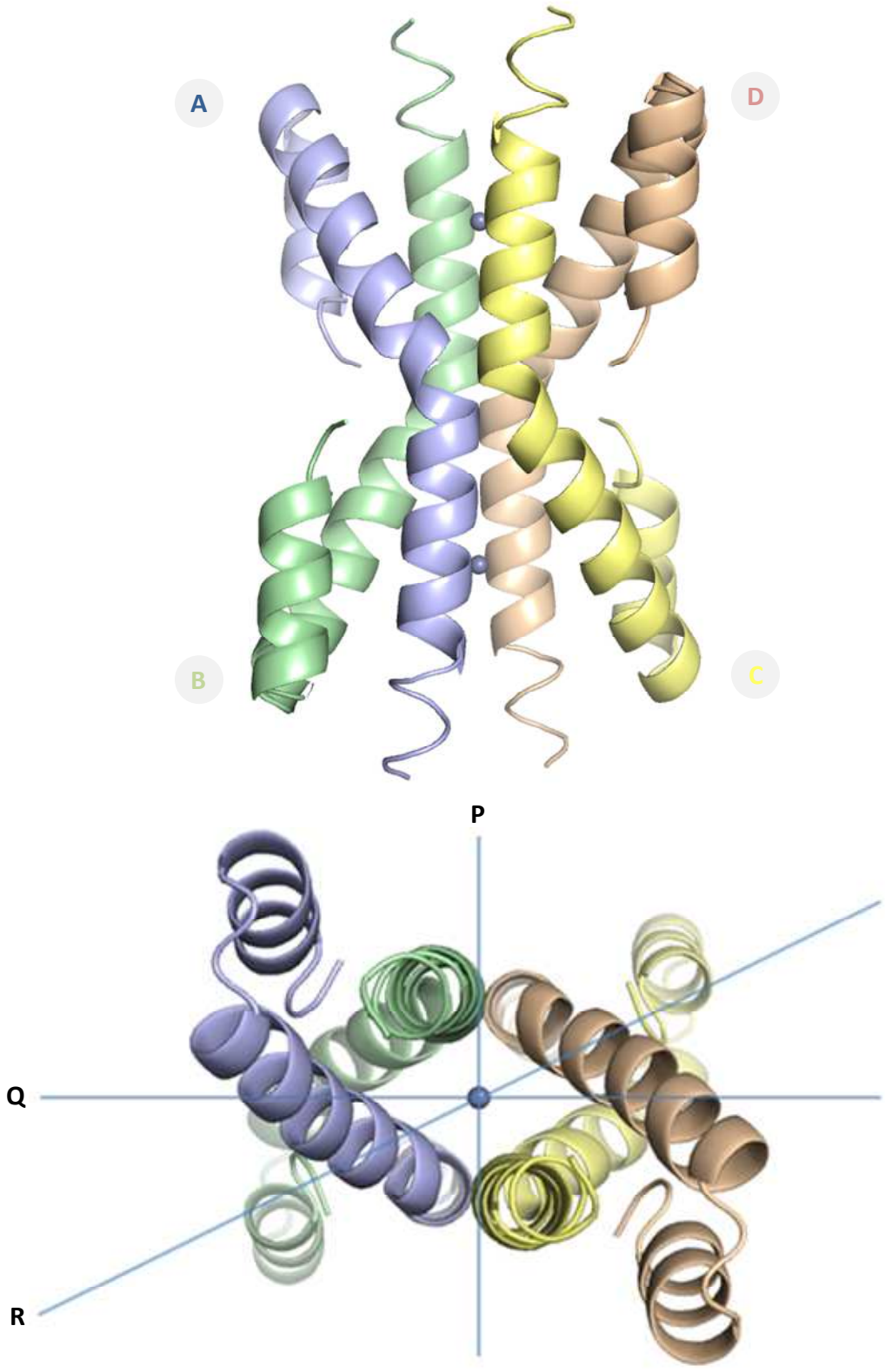


Figure 7.16 Cartoon representation of the quaternary structure of BPSS0211. The two dimers are formed between monomers, A (blue) and B (green), and C (yellow) and D (pink), while the tetramer is formed between the two dimers and contains one metal ion (grey spheres) per dimer at the interface.

The relevant interfaces can be described using the tetramer PQR axes nomenclature, with the Q axis relating the monomers comprising an individual dimer (A to B), and the P and R axes relating the monomer interactions that form the tetrameric protein (A to C and A to D, respectively) (figure 7.16).

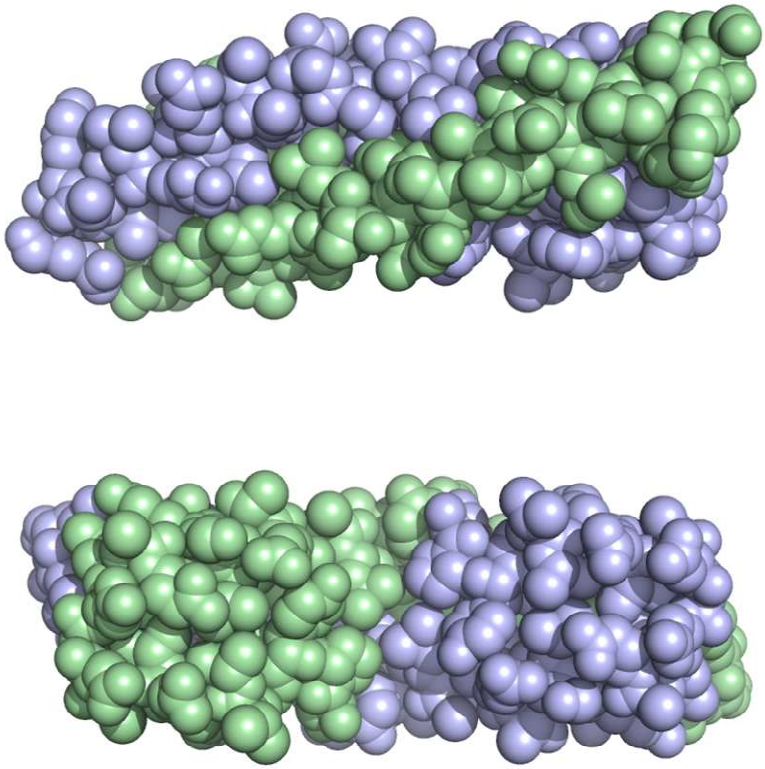
7.9.1 Monomer-monomer interface forming dimeric BPSS0211

The dimer interface between the two monomers (figure 7.17 a), Q-axis, has a buried surface area of 1450 $Å^2$ from each subunit (rounded to the nearest 50 $Å^2$) involving 32 residues providing van der Waals interactions (table 7.6 a) (figure 7.18) and six hydrogen bonds (table 7.6 d) including two salt bridges (table 7.6 e) (figure 7.19) (table 7.6). Analysis using PISA suggested that the extensive level of interaction, involving 49 % of solvent accessible surfaces, between the monomers demonstrates this is a biologically relevant dimer potentially representing the native state of the protein. Furthermore, calculating the likelihood the interactions are crystal contacts, based on the properties of other crystals in the PDB [164], gave a probability of 4.9×10^{-5} that the interaction is insignificant strongly suggesting the dimer interface is biologically relevant.

7.9.2 Monomer-monomer interfaces forming tetrameric BPSS0211

The tetramer interface between the two dimers (figure 7.17 b) is divided into two sets of interactions between a monomer of one dimer and the two monomers of the other dimer around the P and R axes. Together the interactions occur over an area of 400 $Å^2$ for each subunit involving 9 residues providing van der Waals interactions (tables 7.6 b and c) and two possible hydrogen bonds (although these are long) (table 7.6 f) (figure 7.20). The tetramer interactions also include two metal ions, assumed to be zinc, on the 2-fold unit cell axis. Each zinc ion is co-ordinated by the OE2 oxygen of Glu-53 from one member of each dimer across the R-axis (either A and D or B and C) (figure 7.21), with the other two co-ordination sites unoccupied due to problems interpreting the density (section 7.7.3). Analysis using PISA, calculated a total buried surface area, accounting for 77 % of solvent accessible surfaces (28 % of which equates to the tetramer interface), suggesting that the level of interaction between the dimers suggests this is a biologically relevant tetramer. Calculating the likelihood the interactions are crystal contacts [164] gave a probability of 1.9×10^{-2} that the interaction is insignificant suggesting the interface is biologically relevant, although the level of significance is substantially less than for the dimer interface.

(a) Dimeric BPSS0211



(b) Trameric BPSS0211

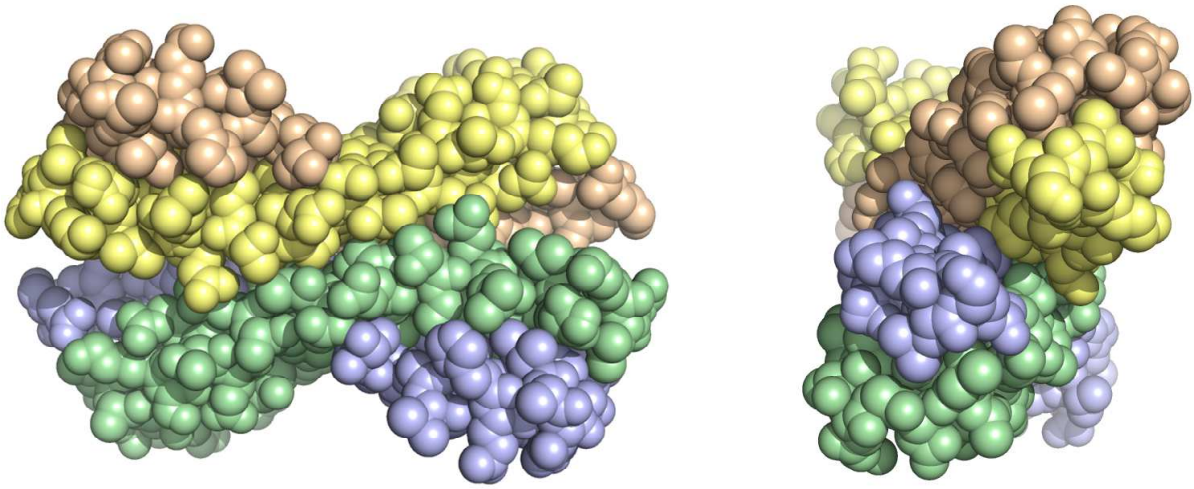


Figure 7.17 Space-filling models of the quaternary structures of BPSS0211. **a** Dimeric BPSS0211. **b** Tertrameric BPSS0211. Each figure has two images taken 90º apart, for **a** this is about the R-axis and **b** about the Q-axis.

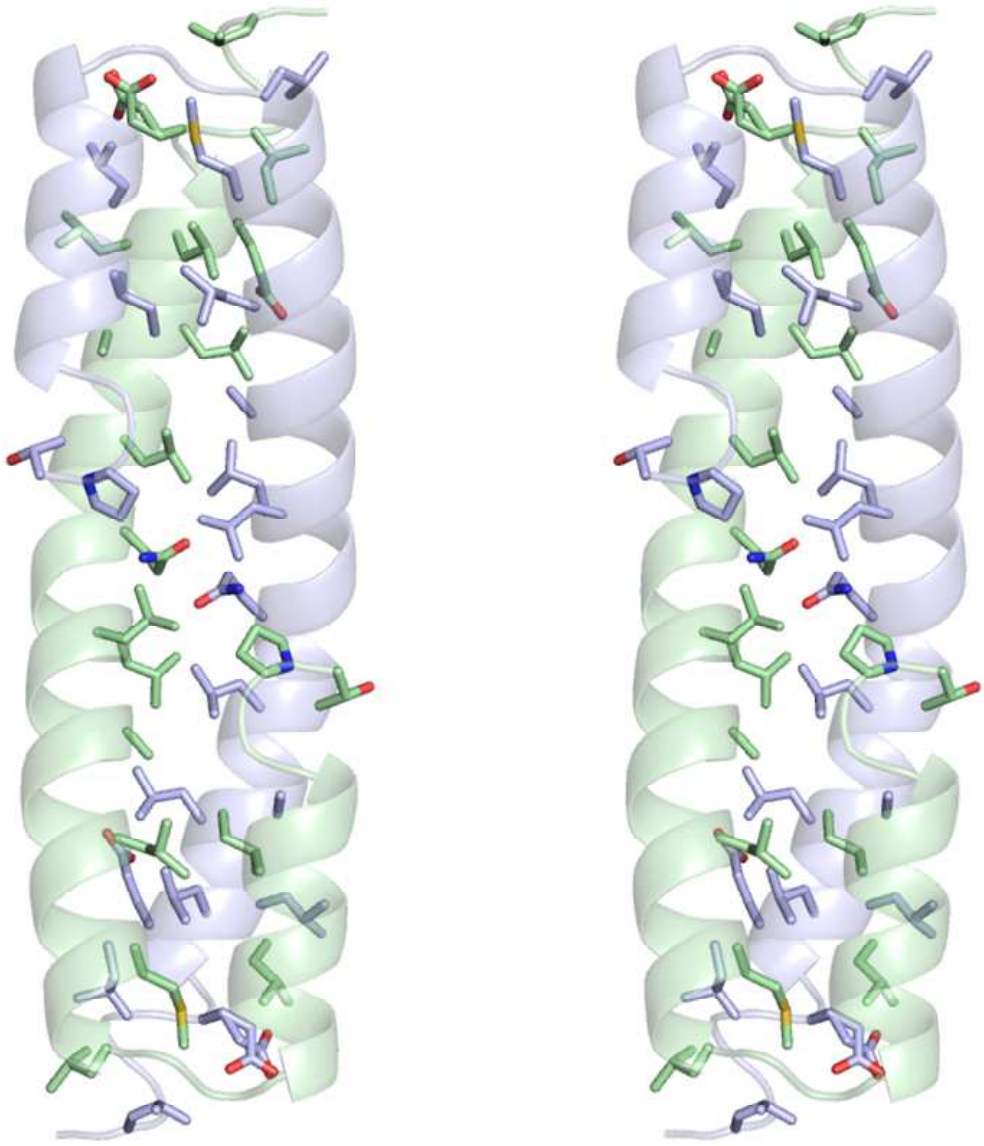


Figure 7.18 Residues involved in van der Waals interactions on the monomer-monomer interface forming dimeric BPSS0211. Residues displayed are those that were predicted to be more than 30 % surface buried.

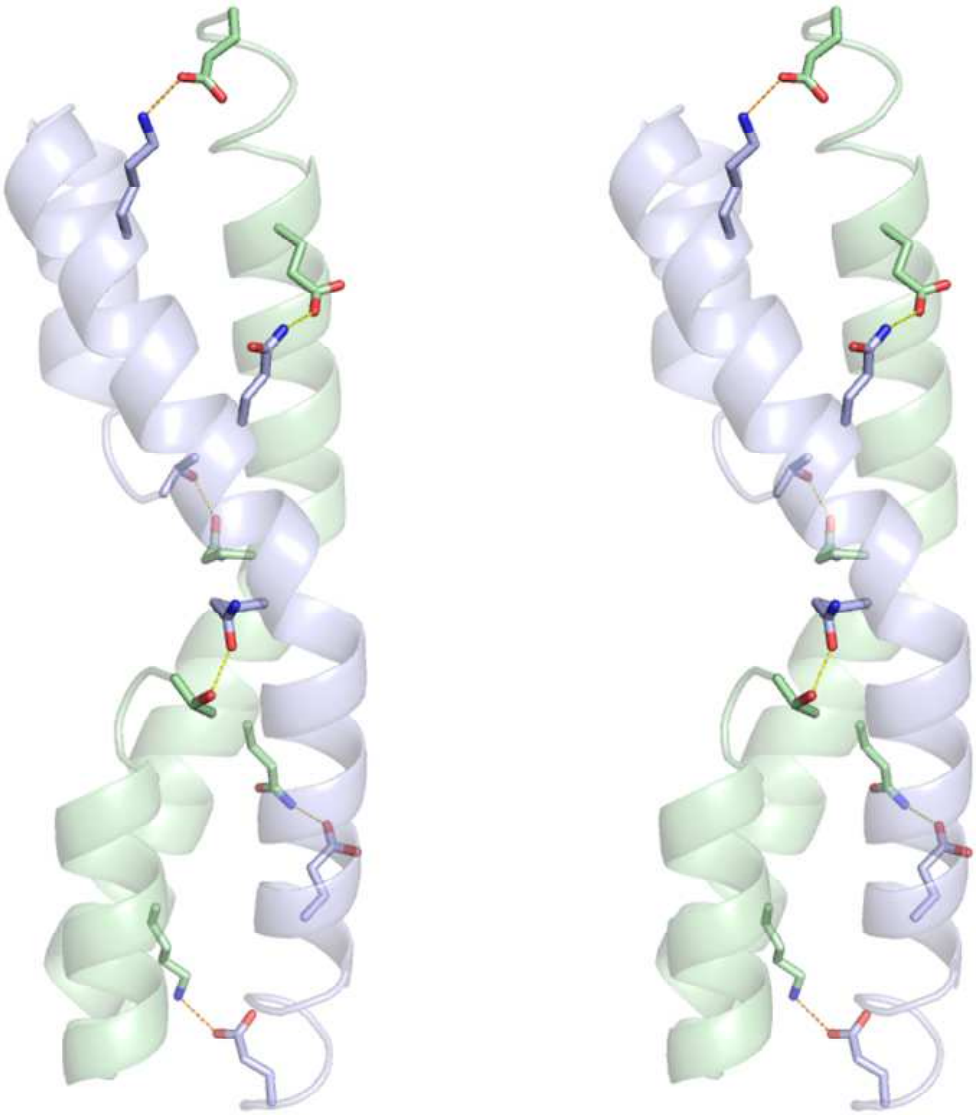


Figure 7.19 Residues involved in polar interactions on the monomer-monomer interface forming dimeric BPSS0211. Hydrogen bonds are shown in yellow and the salt bridges are shown in orange.

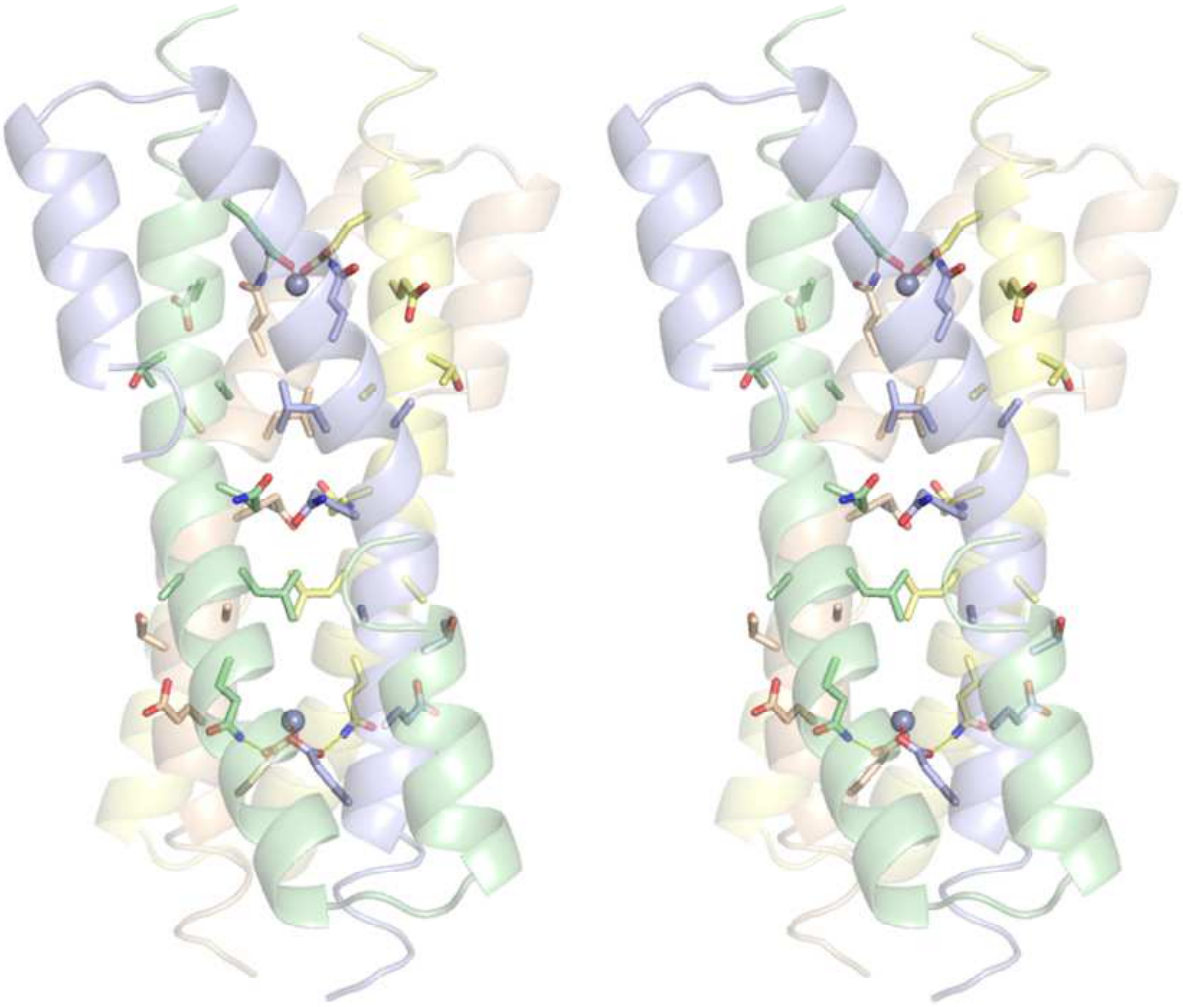


Figure 7.20 Residues involved in the monomer-monomer interfaces forming tetrameric BPSS0211. Residues displayed are those that were predicted to be more than 30 % surface buried including those forming a potential hydrogen bond, Gln-36 and Glu-53, which is shown in yellow. The interactions involving the two zinc ions (grey spheres) are also shown with co-ordination bonds shown in red.

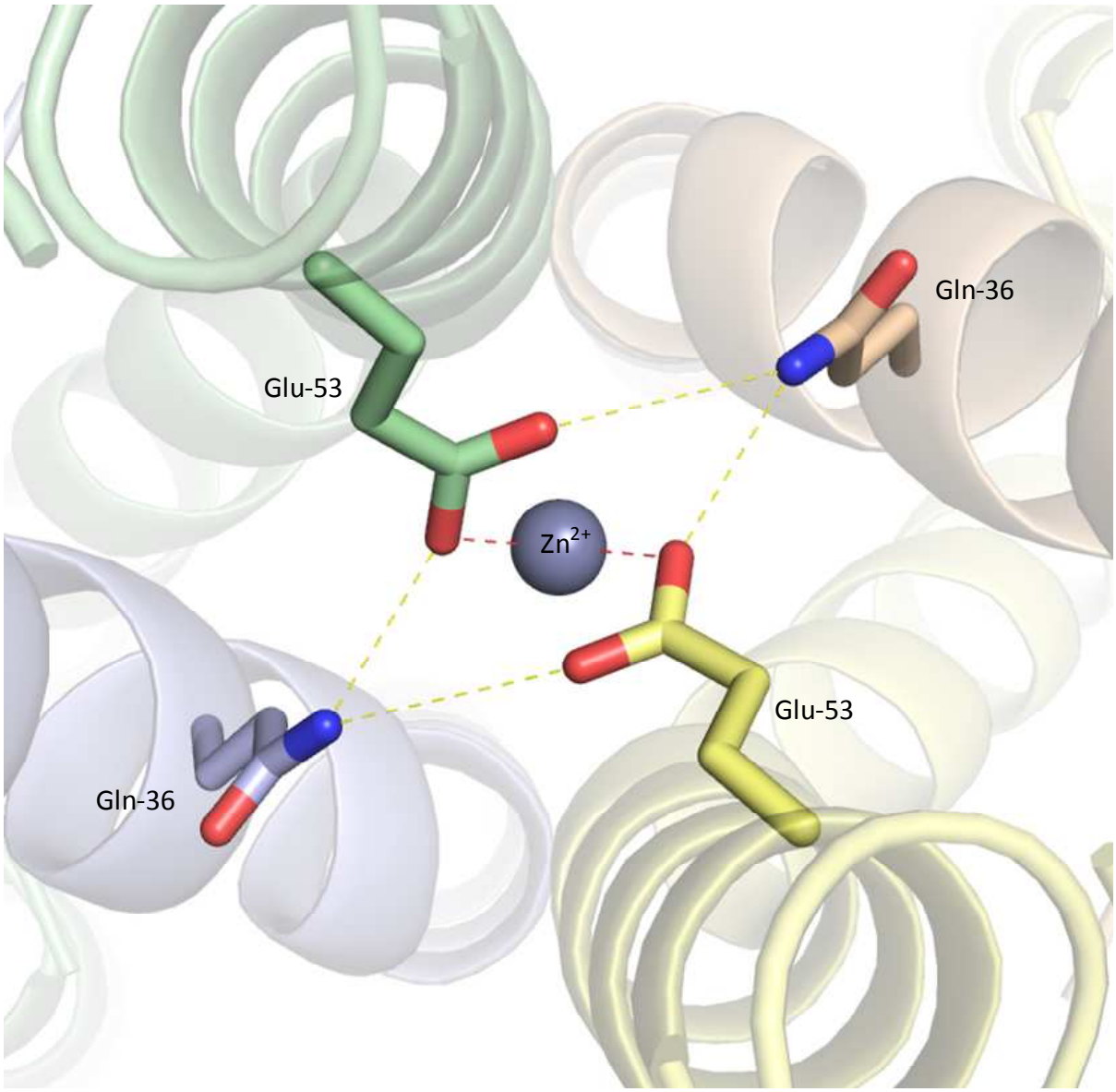


Figure 7.21 Co-ordination of a zinc ion on the tetramer interface of BPSS0211. As indicated in the text, co-ordination of the zinc ion also involves unmodelled molecules from the solvent. The hydrogen bonds between Gln-36 and Glu-53 are shown in yellow and the zinc co-ordination bonds are shown in red.

(a) Surface accessibility of residues on the Q-axis (dimer) interface

Residue	Accessible surface area (Å^2)	Buried surface area (Å^2)	Surface buried (%)
Thr-10	152	76	50
Pro-11	91	56	62
Tyr-12	74	2	2
Gly-13	14	5	32
Val-14	125	0	0
Ala-15	35	0	0
Ile-16	54	54	100
His-17	135	30	22
Gln-18	109	0	0
Ala-19	3	0	10
Ile-20	105	76	73
Ala-21	89	0	0
Asp-22	103	0	0
Gly-23	52	10	20
Asp-24	67	0	0
Leu-25	107	58	54
Ala-26	78	0	0
Gln-27	73	0	0
Met-28	60	60	100
Lys-29	97	28	29
Ser-30	48	0	0
Leu-31	9	2	21
Arg-32	125	60	48
Thr-33	84	0	0
Gln-34	74	0	0
Ala-35	24	24	100
Gln-36	105	29	27
Ala-37	61	0	0
Leu-38	69	43	62
Leu-39	130	80	61
Ala-40	79	0	0
Gln-41	92	40	44
Gln-42	65	47	72
Gly-43	50	0	0
Asn-44	91	15	17
Leu-45	107	107	100
Ala-46	57	14	25
Thr-47	82	0	0
Ala-48	53	45	85
Leu-49	84	74	88
Glu-50	110	0	0
Leu-51	135	49	36
Leu-52	96	95	100
Glu-53	98	25	25
Val-54	81	0	0
Glu-55a+b	86	42	49
Ile-56	90	83	92
Ala-57	66	0	0
Lys-58	144	0	0
Leu-59	144	63	44
Glu-60	178	38	22

(b) Surface accessibility of residues on the P-axis (tetramer) interface

Residue	Accessible surface area (Å^2)	Buried surface area (Å^2)	Surface buried (%)
Thr-10	152	0	0
Pro-11	91	0	0
Tyr-12	74	0	0
Gly-13	14	0	0
Val-14	125	0	0
Ala-15	35	0	0
Ile-16	54	0	0
His-17	135	0	0
Gln-18	109	0	0
Ala-19	3	0	0
Ile-20	105	0	0
Ala-21	89	0	0
Asp-22	103	0	0
Gly-23	52	0	0
Asp-24	67	0	0
Leu-25	107	0	0
Ala-26	78	0	0
Gln-27	73	0	0
Met-28	60	0	0
Lys-29	97	0	0
Ser-30	48	0	0
Leu-31	9	0	0
Arg-32	125	0	0
Thr-33	84	0	0
Gln-34	74	0	0
Ala-35	24	0	0
Gln-36	105	58	55
Ala-37	61	0	0
Leu-38	69	0	0
Leu-39	130	26	20
Ala-40	79	56	71
Gln-41	92	0	0
Gln-42	65	13	20
Gly-43	50	46	92
Asn-44	91	0	0
Leu-45	107	0	0
Ala-46	57	35	62
Thr-47	82	22	26
Ala-48	53	0	0
Leu-49	84	0	0
Glu-50	110	45	41
Leu-51	135	0	0
Leu-52	96	0	0
Glu-53	98	15	15
Val-54	81	0	0
Glu-55a+b	86	0	0
Ile-56	90	0	0
Ala-57	66	0	0
Lys-58	144	0	0
Leu-59	144	0	0
Glu-60	178	0	0

(c) Surface accessibility of residues on the R-axis (tetramer) interface

Residue	Accessible surface area (Å^2)	Buried surface area (Å^2)	Surface buried (%)
Thr-10	152	0	0
Pro-11	91	0	0
Tyr-12	74	0	0
Gly-13	14	0	0
Val-14	125	0	0
Ala-15	35	0	0
Ile-16	54	0	0
His-17	135	0	0
Gln-18	109	0	0
Ala-19	3	0	0
Ile-20	105	0	0
Ala-21	89	0	0
Asp-22	103	0	0
Gly-23	52	0	0
Asp-24	67	0	0
Leu-25	107	0	0
Ala-26	78	0	0
Gln-27	73	0	0
Met-28	60	0	0
Lys-29	97	0	0
Ser-30	48	0	0
Leu-31	9	0	0
Arg-32	125	0	0
Thr-33	84	0	0
Gln-34	74	0	0
Ala-35	24	0	0
Gln-36	105	0	0
Ala-37	61	0	0
Leu-38	69	0	0
Leu-39	130	23	18
Ala-40	79	0	0
Gln-41	92	0	0
Gln-42	65	0	0
Gly-43	50	0	0
Asn-44	91	0	0
Leu-45	107	0	0
Ala-46	57	0	0
Thr-47	82	0	0
Ala-48	53	0	0
Leu-49	84	0	0
Glu-50	110	0	0
Leu-51	135	0	0
Leu-52	96	0	0
Glu-53	98	36	37
Val-54	81	0	0
Glu-55a+b	86	0	0
Ile-56	90	0	0
Ala-57	66	0	0
Lys-58	144	0	0
Leu-59	144	0	0
Glu-60	178	0	0

(d) Unique hydrogen bonds across the Q-axis

Residue in monomer 1	Residue in monomer 2	Bond distance (Å)
Thr-10 [OG1]	Gln-41 [OE3]	2.54
Lys29 [NZ]	Glu-60 [OE2]	2.89
Gln-36 [NE2]	Glu-53 [OE2]	3.16

(e) Unique salt bridges across the Q-axis

Residue in monomer 1	Residue in monomer 2	Bond distance (Å)
Lys-29 [NZ]	Glu-60 [OE2]	2.89

(f) Unique hydrogen bonds across the P-axis

Residue in monomer 1	Residue in monomer 2	Bond distance (Å)
Gln-36 [NE2]	Glu-53 [OE1]	3.85

(g) Overall statistics

Interface axis	Q	P	R
Symmetry operator	-x+y+1,y,-z+1	-x+1,-y,z	x-y,x-y,z-1/3
Interface area ($Å^2$)	1429.0	315.8	59.2
Number of interfacing residues from each monomer	32	9	2
Number of hydrogen bonds from each monomer	6	2	0
Number of salt bridges from each monomer	2	0	0

Table 7.6 Dimer interfaces of BPSS0211 around the P, Q and R axes.

7.10 Residue conservation across BPSS0211, BPSS0212, BPSS0213 and homologs from other species

A Blast search [165] was conducted using the BPSS0211 protein sequence to identify all related proteins. The results fall into two sets of two categories. The first is the length of the construct, whether like BPSS0211 the domain exists in isolation or like BPSS0212 and BPSS0213 it is part of a larger construct. The pattern of conservation between the two forms is largely similar within organisms (data not shown). The second characteristic is defined by the species the protein derived from, those from members of the pseudomallei group have sequence identities of over 80 % and those from other *Burkholderia* species and more distantly related organisms have identities between 62 and 36 %. Residues that were conserved in over 90 % of the aligned sequences were identified using Clustal Omega [166] (figure 7.22) and mapped onto the protein structure. None of the conserved residues lie on the P or R axes tetramer interfaces but are located either on the Q-axis dimer interface or at the interface of helices I and II in the monomer (figures 7.23 and 7.24).

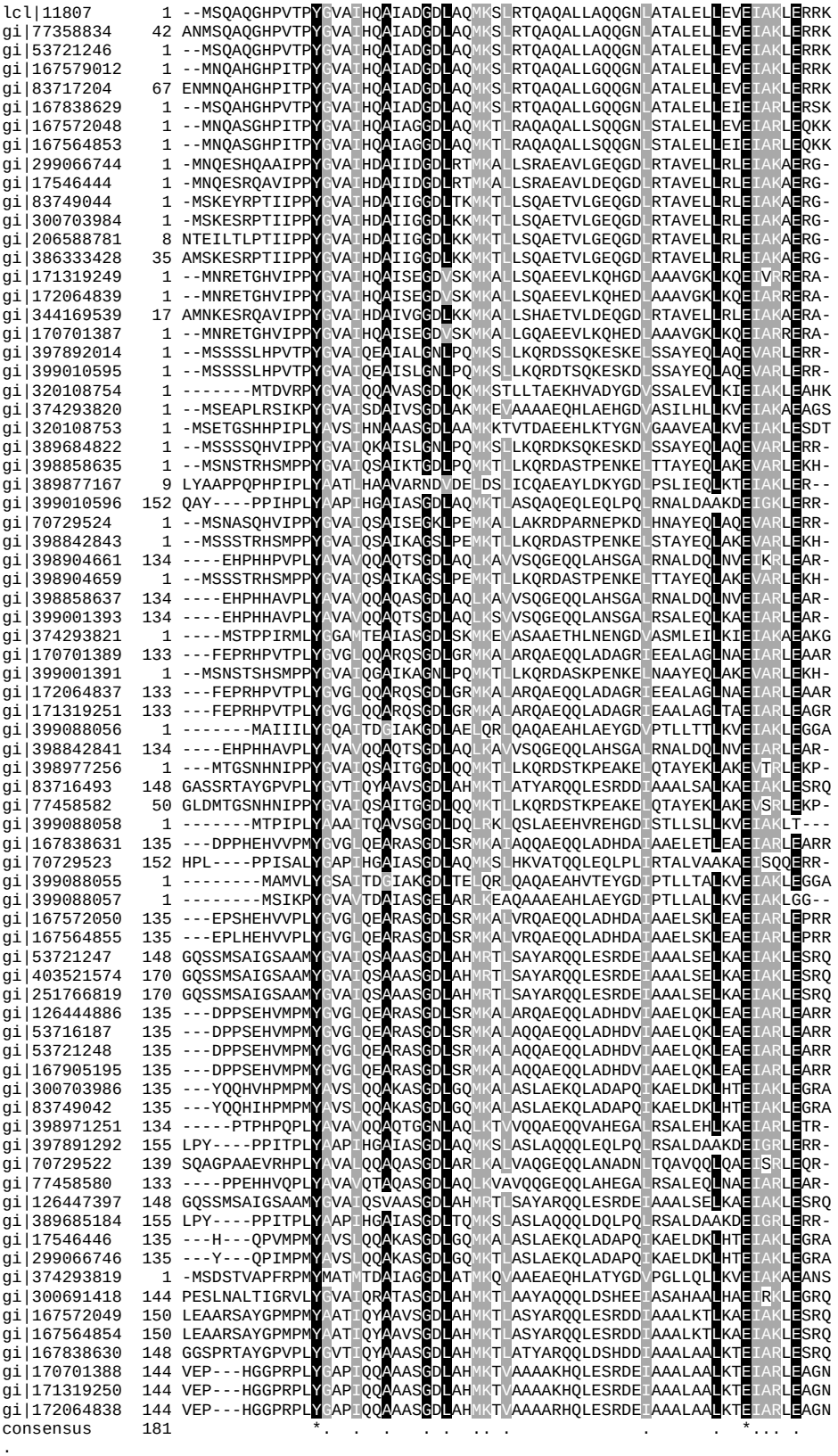


Figure 7.22 Residues conserved in over 90 % of BPSS0211 homologs. The alignment has been produced using Clustal Omega [162] and BOXSHADE [135] and the consensus is shown. Residues that are completely conserved are shaded black and shown as (*) in the consensus, more than 95 % are also shaded black but shown as (.), more than 90 % are shaded grey and shown as (.) in the consensus, with the remaining residues left blank.

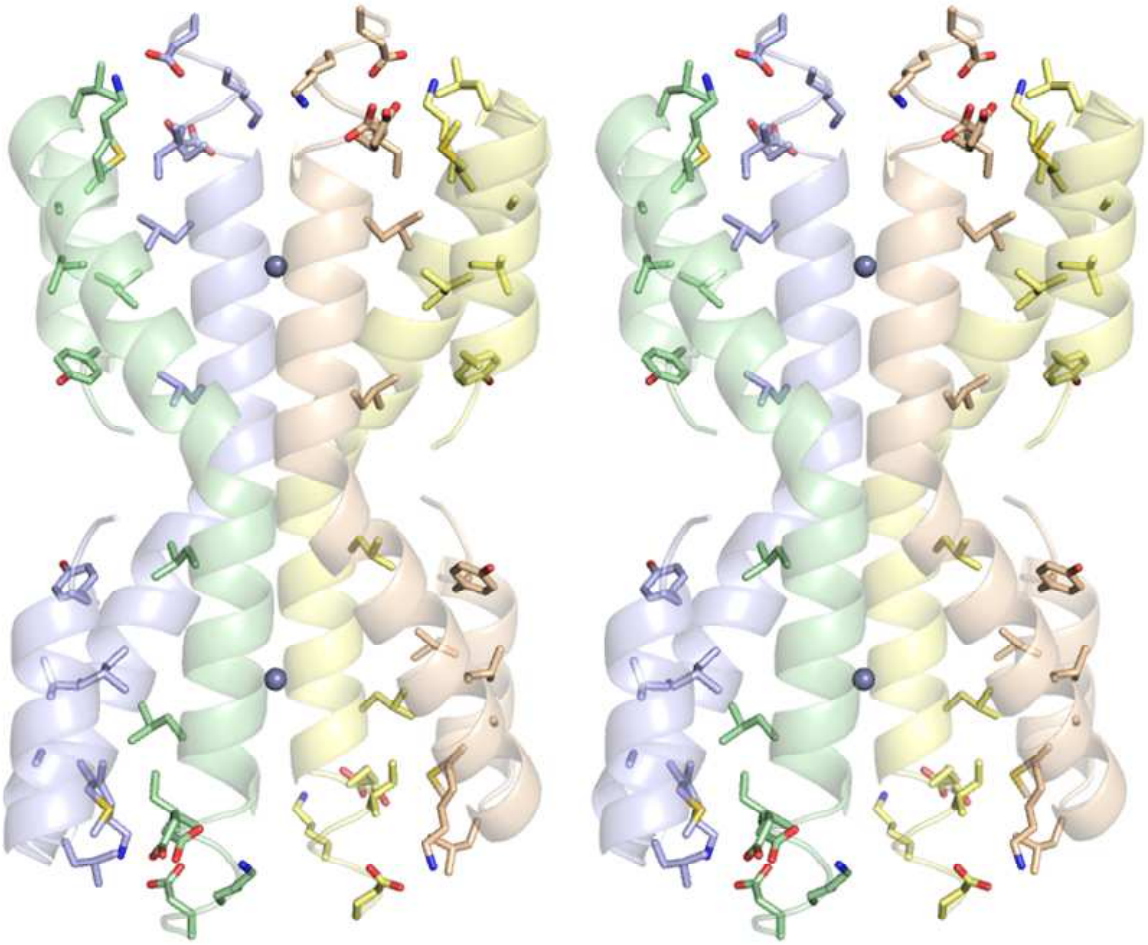


Figure 7.23 Conserved residues in BPSS0211 are located on the dimer interface or at the interface of helices I and II in the monomer.

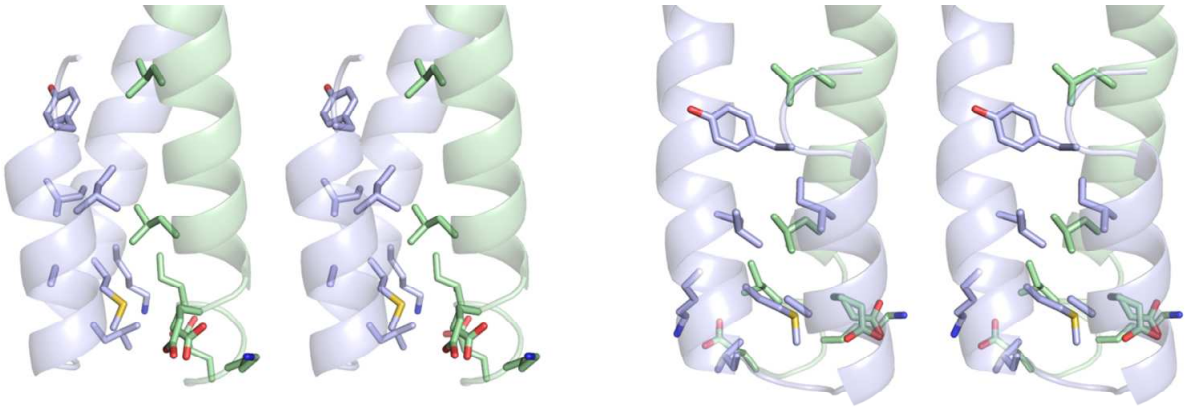


Figure 7.24 The unique residues conserved in a dimer of BPSS0211. The two stereo images are taken 90º apart and show the residues positioned at the dimer interface or the interface of helices I and II in the monomer.

7.11 BPSS0211 represents an oligomerisation domain

The quaternary structure of BPSS0211 suggests that it represents an oligomerisation domain that can exist on its own (BPSS0211) or as part of a larger construct (BPSS0212 and BPSS0213). The pattern of residue conservation between BPSS0211, BPSS0212 and BPSS0213, within *Burkholderia pseudomallei* and homologs from other species, show those involved in dimer formation, but not tetramer formation are heavily conserved. Coupled with the quaternary structure, this suggests that BPSS0211 represents an oligomerisation platform which mediates assembly of a complex, whose stoichiometry between combinations of BPSS0211, BPSS0212 and BPSS0213, remains to be determined.

8.0 Studies on the protein BPSS0212

This section covers the purification, crystallisation, data collection and attempts to solve the structure of the target BPSS0212.

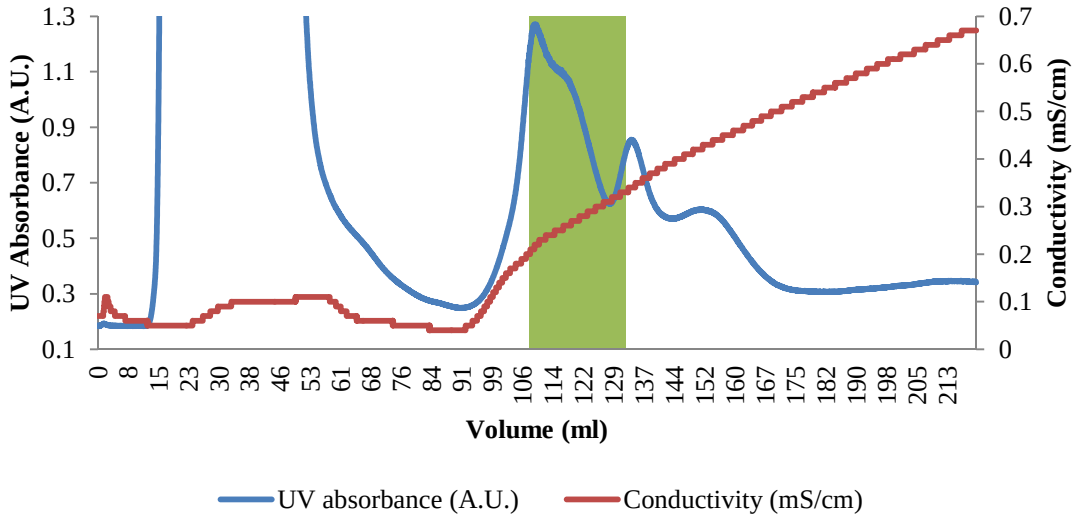
8.1 Protein purification for BPSS0212

8.1.1 Native protein purification

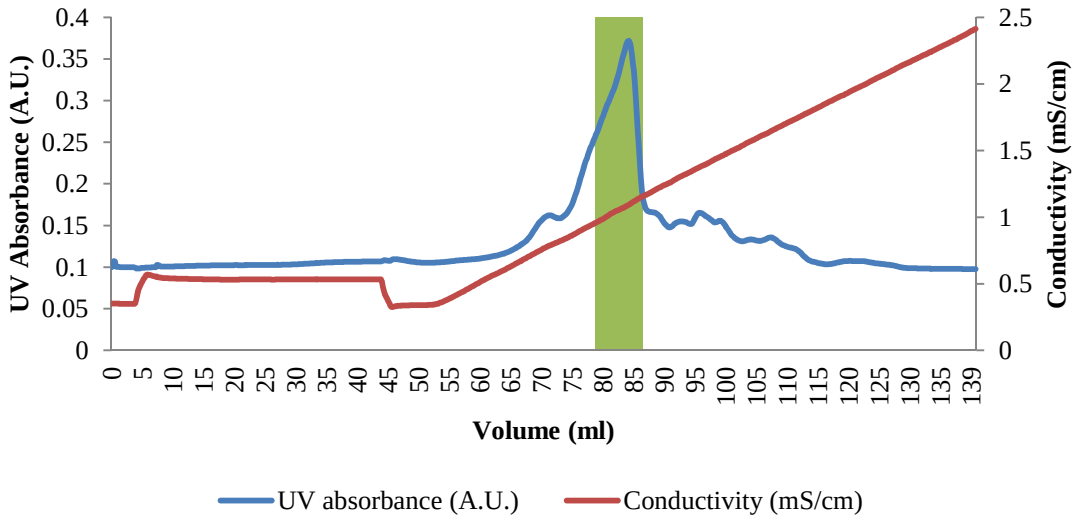
Approximately 3 g of cell paste was resuspended in 30 ml 50 mM TRIS pH 8.0 and disrupted by sonication. Cell debris and insoluble protein was removed by centrifugation at 70,000 g for 15 minutes and the supernatant was loaded onto a DEAE-Sepharose fast flow column equilabrated with 50 mM TRIS pH 8.0. A 200 ml gradient from 0 to 500 mM NaCl was then applied to the column and 8 ml fractions were collected (figure 8.1 a). Fractions were analysed by SDS-PAGE (figure 8.2) and those containing BPSS0212 were combined. The sample was diluted two-fold before loading onto a ResourceQ column equilibrated with 50 mM TRIS pH 6.3. A 120 ml gradient was applied to the column and 2.5 ml fractions were collected (figure 8.1 b). Peak fractions were combined and concentrated to 1.5 ml using a Vivaspin concentrator with a 10 kDa MWCO. This was then loaded onto a 1.6 x 60 cm Superose 6 gel filtration column equilibrated in 50 mM TRIS pH 8.0 and 500 mM NaCl and eulted using the same buffer collecting 2 ml fractions (figure 8.1 c). Peak fractions were combined, the buffer was exchanged for 10 mM TRIS pH 8.0 and the protein was concentrated to 5 mg ml^{-1} for use in crystallisation trials using the same Vivaspin concentrator. The overall yield of protein was 7 mg which was estimated by SDS-PAGE to be over 90 % pure (figure 8.2)

Figure 8.1 Chromatogram traces for the purification of BPSS0212. **a** DEAE purification step showing column loading and elution. 8 ml fractions were collected starting at the beginning of the gradient at 36 ml. **b** ResourceQ purification step showing column loading and elution. 2 ml fractions were collected starting at the beginning of the gradient at 39 ml. **c** Gel filtration purification step showing elution with 2 ml fractions collected throughout. For all traces, green highlighted regions indicate volumes taken for subsequent purification steps or as pure protein.

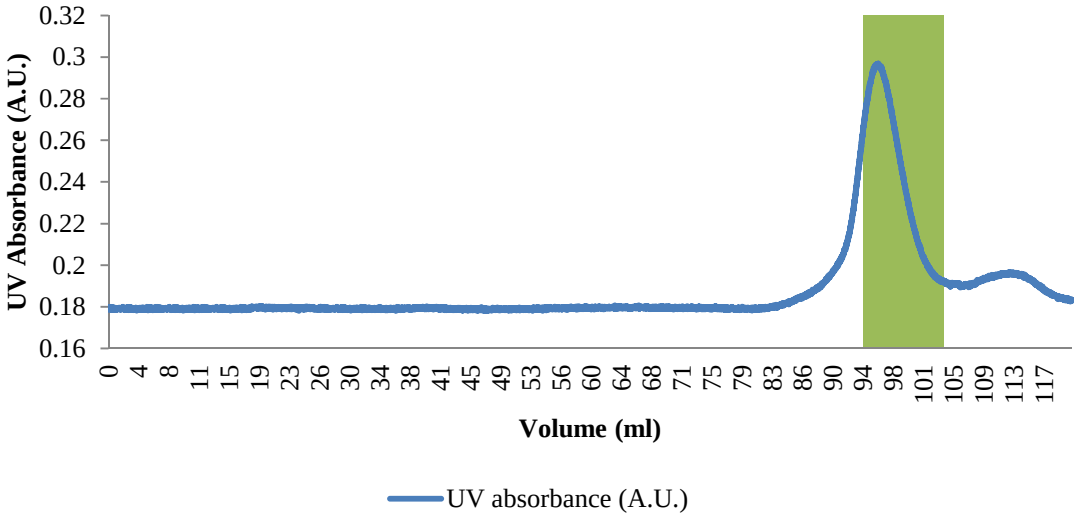
(a) BPSS0212 DEAE purification step



(b) BPSS0212 ResourceQ purification step



(c) BPSS0212 Superose 6 purification step



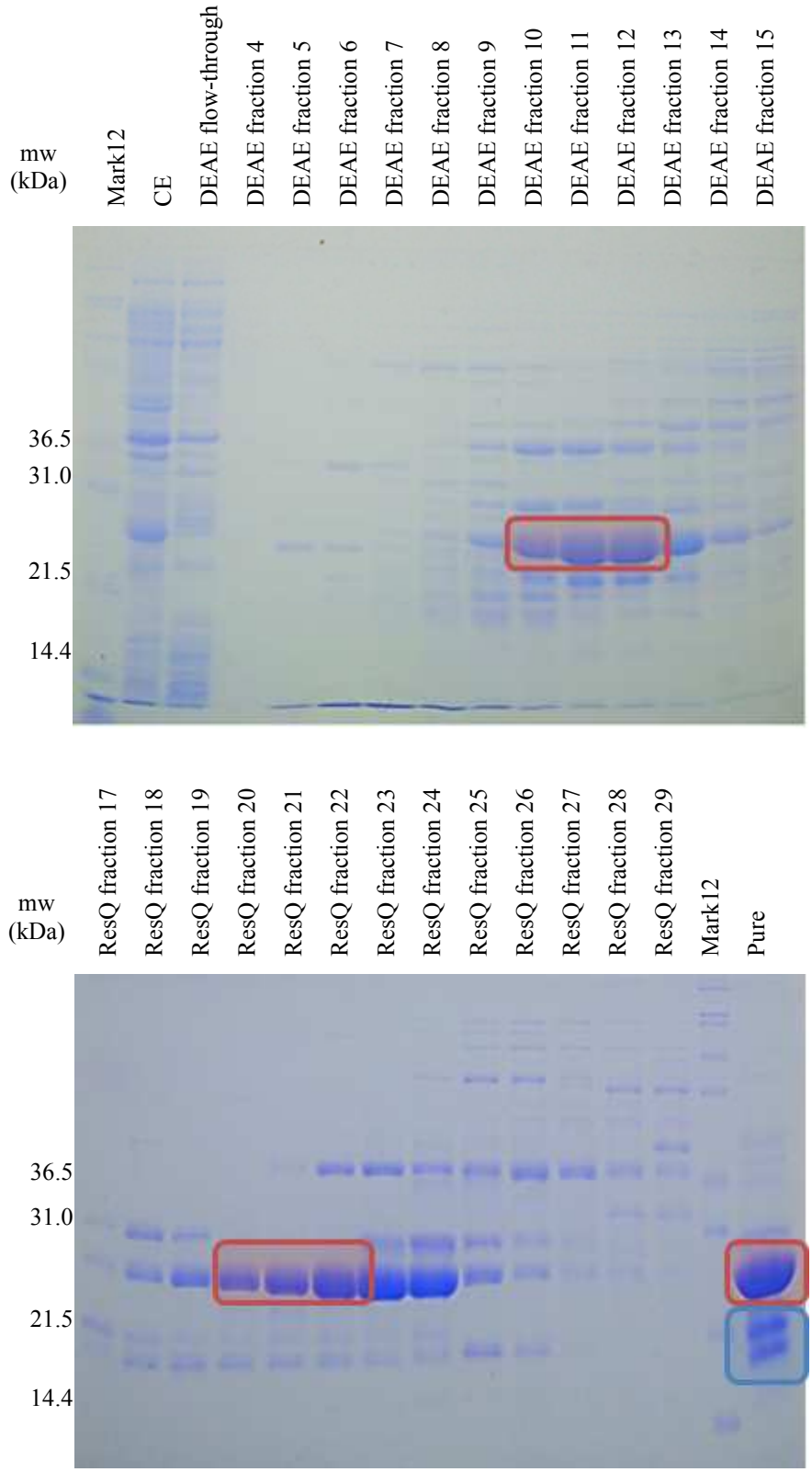


Figure 8.2 SDS-PAGE gel showing the purification of BPSS0212. The molecular weight of BPSS0212 is 22.5 kDa and the red highlighted bands indicate the protein in fractions taken for subsequent purification steps or as pure protein. The blue highlighted bands indicate degradation products of BPSS0212.

8.1.2 Seleno-methionine protein purification

Protein containing seleno-L-methionine was purified using the same techniques as for native protein. Approximately 3 g of cell paste was used resulting in a yield of 5 mg protein at 5 mg ml^{-1} for use in crystallisation trials.

8.1.3 Purification analysis

The elution profile for BPSS0212 from the gel filtration column shows a broad peak roughly corresponding to approximately 35 kDa, a molecular weight best explained by a dimeric form of the protein (figure 8.1 c). SDS-PAGE analysis of the purification shows the sample for crystallisation studies was over 90 % pure but suffering severely from degradation even during the timescale of the purification process (figure 8.2). A sample kept at 4 ºC for two weeks following purification was sent for mass spectrometry (Simon Thorpe, University of Sheffield) to identify the degradation products (figure 8.3). The results showed several fragments of different molecular weights although none corresponded perfectly with any particular fragment of the protein.

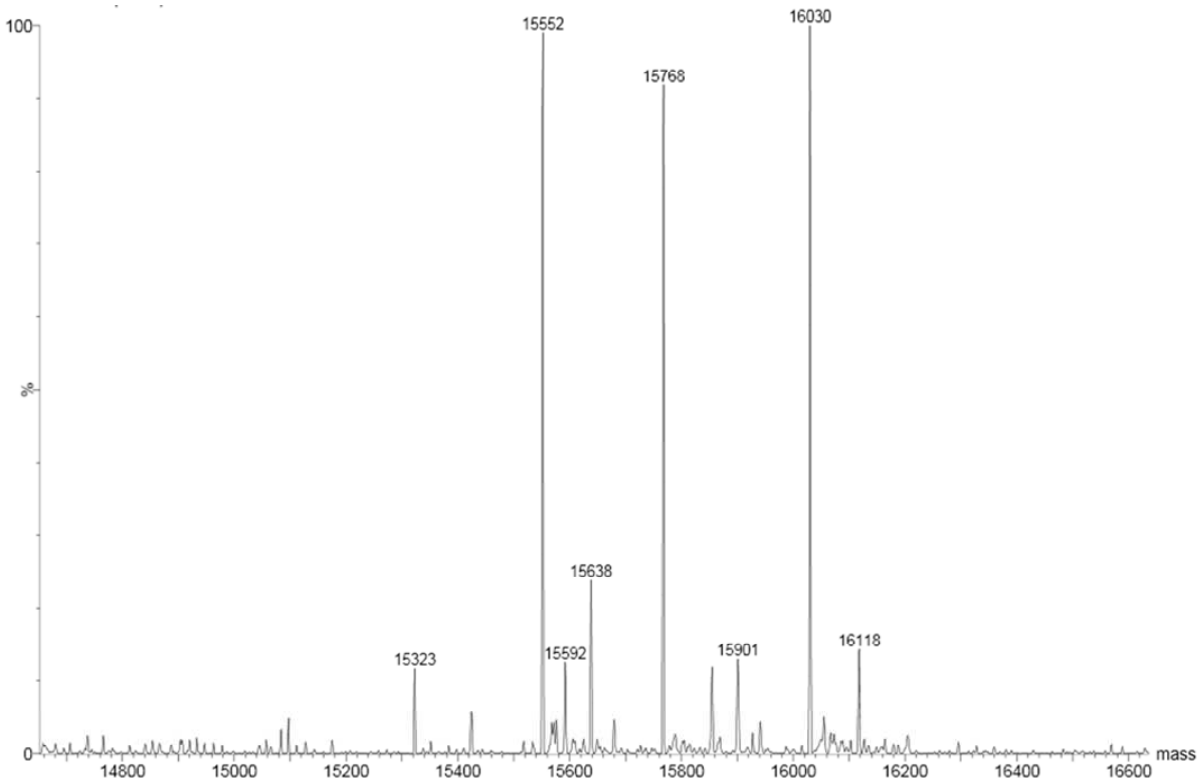


Figure 8.3 Mass spectrometry results for a sample of purified BPSS0212.

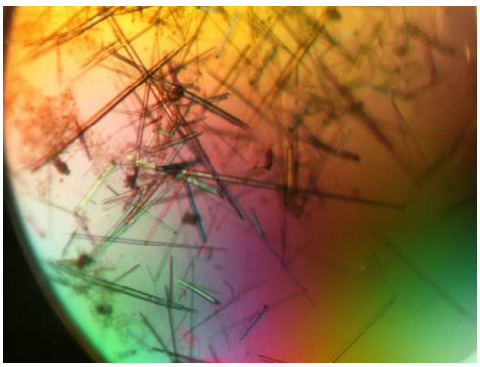
8.2 Protein crystallisation for BPSS0212

8.2.1 Native protein crystallisation

Seven initial 96 condition robot screens, the Ammonium sulphate, Classics, JCSG+, MPDs, PACT, PEGs and pH clear suites, were conducted using purified BPSS0212 at 5 mg ml^{-1} in 10 mM TRIS pH 8.0 and 100 mM NaCl. The screens were conducted using 200 nl of protein mixed with 200 nl of well solution and the trays were incubated at 17 ºC. Over fifty hits were obtained in many different conditions producing mainly needle clusters or rod shaped crystals. Some of these conditions were selected for optimisation trials however all attempts to optimise any of the initial hits were unsuccessful with either no crystals or delicate thin needle clusters developing that were unsuitable for data collection.

8.2.2 Seleno-methionine protein crystallisation

Four initial 96 condition robot screens, the Classics, JCSG+, PACT and Pegs, were conducted using purified seleno-methionine BPSS0212 at 5 mg ml^{-1} in 10 mM TRIS pH 8.0 and 100 mM NaCl. The screens were conducted using 200 nl of protein mixed with 200 nl of well solution and the trays were incubated at 17 ºC. Similar hits were obtained to the native protein, appearing in many different conditions and similarly to the native protein these crystals defied optimisation trials.

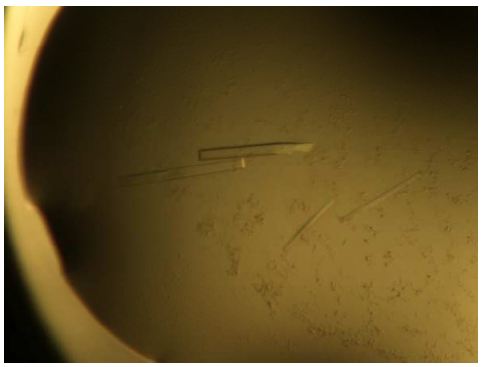


Native – Pegs D12 – robot screen

100 mM TRIS pH 8.5

15 % (w/v) PEG 20,000

Drop size 200 nl protein + 200 nl well solution



Se-met – Pegs D12 – robot screen

100 mM TRIS pH 8.5

15 % (w/v) PEG 20,000

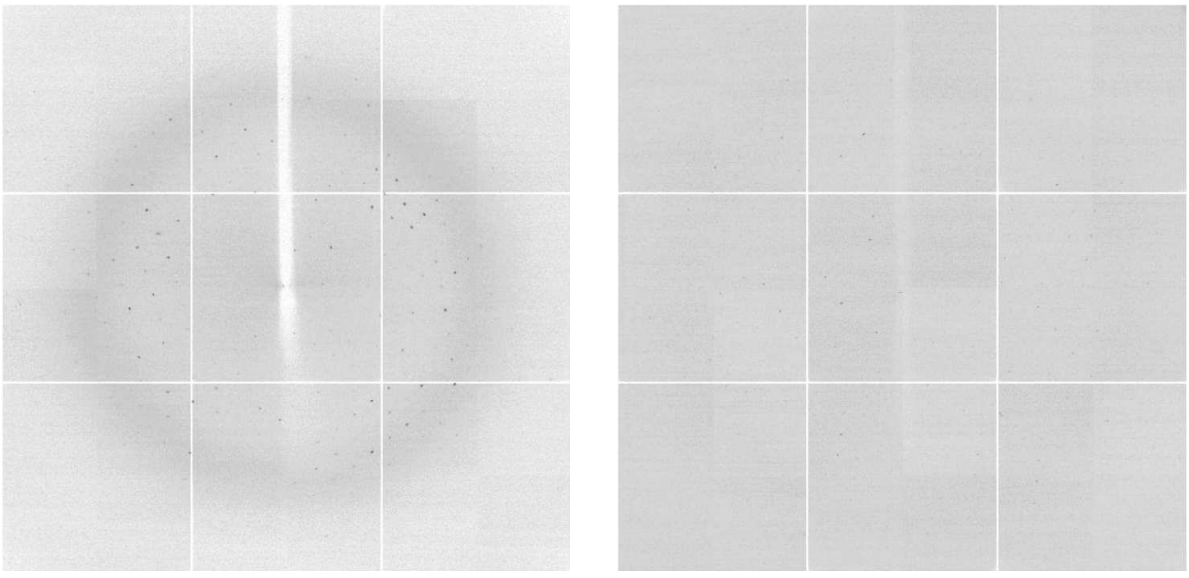
Drop size 200 nl protein + 200 nl well solution

Figure 8.4 Photographs of BPSS0212 native and seleno-methionine protein crystals.

8.3 BPSS0212 native data

8.3.1 Native data collection

Native crystals were selected for data collection from conditions in the robot trials which produced the largest best defined crystals. They were looped and placed into a cryoprotectant solution consisting of the well solution with the addition of 30 % (v/v) ethylene glycol. They were then looped again and flash frozen in a stream of gaseous nitrogen at 100 ºK and mounted onto the detector. Initial diffraction analysis was conducted in order to determine the diffraction quality of the crystals using an in-house Rigaku MM007 rotating anode generator and a MAR 345 research image plate. Two images were collected 90º apart with 1º oscillation. Crystals from conditions Pegs D12 produced reasonable diffraction images with data extending beyond 3.5 Å. These crystals were saved and taken to the I02 beamline of the Diamond light source, Oxford. Two initial test images were taken 90º apart with 1º oscillation to ensure crystal centering and to obtain a collection strategy using the auto-indexing and collection strategy components of Mosflm [146]. For the dataset 210 images were collected with 0.5º oscillation per image at a crystal to detector distance of 378 mm using an ADSC Q315r detector. Data extending to 2.5 Å were collected (figure 8.5).



Native dataset **Seleno-methionine peak dataset**

Figure 8.5 Diffraction images of native and seleno-methionine crystals of BPSS0212.

8.3.2 Native data processing

The X-ray diffraction datasets were auto-processed at Diamond through the xia2 system using the 3dii mode [147]. Data were indexed and integrated by XDS and scaled by XSCALE [148]. Initially the space group was thought to be P422 as described by Mosflm [146] following data collection using the home source and collection of the first two images, however once a complete dataset had been collected the data indexed to a primitive orthorhombic space group, most likely $P2_12_12$ based on the pattern of systematic absences, with the unit cell parameters a = 40.5, b = 97.7 Å, c = 40.0 Å, (table 8.1). As the crystal contents were unknown, Matthews coefficients were calculated using Mattprob [149] for a molecular weight of 16 kDa, decided by analysis of the fragments identified by mass spectrometry (figure 8.3), suggesting a single molecule in the asymmetric unit. The real contents of the crystals remain unknown and the predicted contents were treated with caution (table 8.2).

Dataset	**Native protein crystal**	**Seleno-methionine protein crystal**		
		peak	**inflection**	**high energy remote**
Space group	Orthorhombic P	Orthorhombic P	Orthorhombic P	Orthorhombic P
Unit cell parameters				
a (Å)	40.47	40.22	40.26	40.22
b (Å)	97.74	97.52	97.60	97.55
c (Å)	38.95	38.80	38.82	38.81
α (º)	90.00	90.00	90.00	90.00
β (º)	90.00	90.00	90.00	90.00
γ (º)	90.00	90.00	90.00	90.00
Energy (eV)	12658	12655	12657	13450
Resolution range (Å)	40.47 – 2.30	31.03 – 2.50	27.95 (2.44)	40.22 – 2.54
Unique reflections	26781 (1154)	35269 (1369)	72911 (2477)	34417 (1394)
R_{merge}	0.073 (0.431)	0.097 (0.505)	0.099 (0.731)	0.116 (0.614)
R_{pim}	0.048 (0.323)	0.059 (0.324)	0.032 (0.304)	0.054 (0.376)
Completeness (%)	96.3 (75.2)	98.3 (85.9)	96.9 (80.0)	99.2 (90.6)
Anomalous completeness (%)	91.5 (66.5)	97.9 (83.6)	96.5 (78.6)	99.0 (90.4)
Multiplicity	3.8 (3.0)	6.3 (4.0)	12.4 (7.4)	6.4 (4.2)
Anomalous multiplicity	2.1 (1.6)	3.4 (2.0)	6.9 (3.9)	3.5 (2.1)
Mean (I)/σ(I)	10.8 (2.1)	15.1 (2.2)	22.2 (2.5)	12.4 (2.0)

Table 8.1 Data collection statistics for native and seleno-methionine BPSS0212 crystals. Numbers in parentheses indicate values for the highest resolution shell.

Molecules in the AU	Probability (based data resolution)	Probability (all proteins in the pdb)	Vm (Å^3 / Da)	Solvent content (%)	Molecular weight (Da)
1	1.0000	1.0000	2.47	50.26	16000

Table 8.2 Matthews coefficient calculations and probabilities for native crystals of BPSS0212. The results show a possibility of a single protein molecule inhabiting the asymmetric unit.

8.4 Phasing by molecular replacement for BPSS0212

The structure of BPSS0211 solved as part of this structural genomics project (section 7) was used to create a search model. The model was cut back to a poly alanine chain using chainsaw [150] before use for an automated search for two copies of the molecule in Phaser [151] with the native data set. The data were input and Phaser was run in $P2_12_12$ and all alternative related space groups. Unfortunately the best result was of a poor quality with low Z-scores of 3.6 and 4.4 for the rotation and translation searches respectively. The resulting refined model was also unconvincing, with an R-factor of 0.61, after refinement with REFMAC5 [152], suggesting a solution had not been found.

8.5 BPSS0212 Seleno-methionine data

8.5.1 Seleno-methionine data collection

Seleno-methionine crystals were selected from the same robot screen condition as for native crystals as the crystal contents and form could be assumed to be similar. A number of crystals were mounted using the same methodology as for native crystals and saved. Data was collected from a single crystal using a synchrotron at the I24 microfocus beamline of the Diamond light source, Oxford. A fluorescence scan at the Selenium K edge was conducted to confirm Selenium incorporation and select X-ray energies to collect data for in a MAD experiment (figure 8.6 a). Three energies were selected for data collection based on the fluorescence absorbance spectrum. The peak (maximised f'') and inflection point (minimised f') were selected as 12657 eV and 12655 eV respectively using CHOOCH [157] (figure 8.6 b) and the high energy remote (maximised Δf' from inflection) was selected as 13450 eV. Two initial test images were taken 90º apart with 1º oscillation to ensure crystal centering and to obtain a collection strategy using the auto-indexing and collection strategy components of

Mosflm [146]. For the peak dataset 720 images were collected and for the inflection and remote datasets 360 images were collected with 0.2º oscillation per image at a crystal to detector distance of 422 mm using an ADSC Q315r detector. Data extending to 2.4 Å were collected for each X-ray energy (figure 8.5).

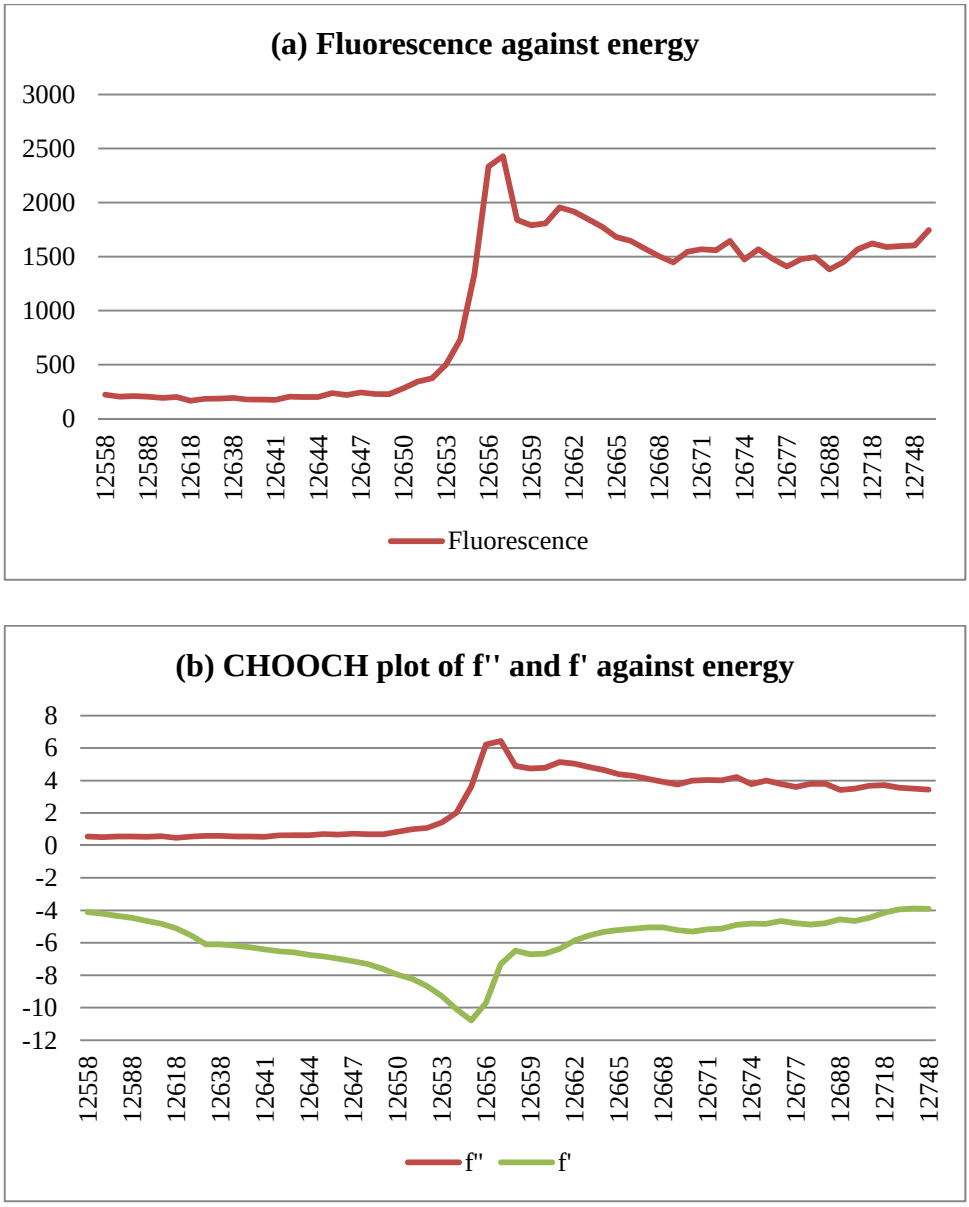


Figure 8.6 Selenium K-edge fluorescence scan and CHOOCH plot for BPSS0212 seleno-methionine crystals. **a** The fluorescence spectrum at the Selenium K-edge confirmed the incorporation of Selenium into the protein crystals and showed maximal fluorescence at 12657 eV, 1 eV below the theoretical value of 12658 eV. **b** Three energies were chosen for the MAD experiment based on the CHOOCH plot, the peak to obtain maximum f'', the inflection point to obtain minimum f' and the high energy remote to maximise Δf'. The three energies used were 12657 eV for the peak, 12655 eV for the inflection and 13450 eV for the high energy remote.

8.5.2 Seleno-methionine data processing

The X-ray diffraction datasets were auto-processed at Diamond through the xia2 system using the 3dii mode [147]. Data were indexed and integrated by XDS and scaled by XSCALE [148]. Data indexed to the same space group as for native data with similar unit cell parameters (table 8.1).

8.6 Experimental phasing for BPSS0212

Initial phasing was conducted with the programmes of the SHELX package [154] using the HKL2MAP graphical user interface for SHELXC and SHELXD [155]. The native protein crystal dataset along with the seleno-methionine protein crystal peak, inflection and high energy remote datasets were inputted into SHELXC. The seleno-methionine datasets contained significant anomalous signal beyond 3.5 Å (figure 8.7). SHELXD was used to calculate the heavy atom substructure using all data to 3.5 Å searching for different numbers of sites in all primitive orthorhombic space groups. The best solution was identified with a correlation coefficient of 53.60 and a Patterson figure of merit of 25.30 searching for 7 Selenium sites in space group $P2_12_12$. A total of 4 good potential heavy atom sites were found (figure 8.8). Preliminary protein phasing, density modification and initial model building were conducted using SHELXE-beta with auto-tracing. Five rounds of twenty cycles of phase calculation and density modification followed by auto-tracing were carried out for both the inverted and the original hands of the Selenium substructure using a predicted solvent content of 50 %. The solutions for the two sub-structure hands gave values of 0.419 and 0.467, and 0.686 and 0.693, for the contrast and connectivity of the original and inverted hands respectively, preventing the assignment of a hand to the substructure based on these figures of merit. SHELXE produced models for the original and inverted hands consisting of 15 and 24 residues respectively, with correlation coefficients for partial structure against the data of 2.60 % and 5.98 %. Inspection of the electron density maps for the two hands provided no further insight into the correct substructure hand, and coupled with the models it was unconvincing an interpretable solution had been reached (figure 8.9).

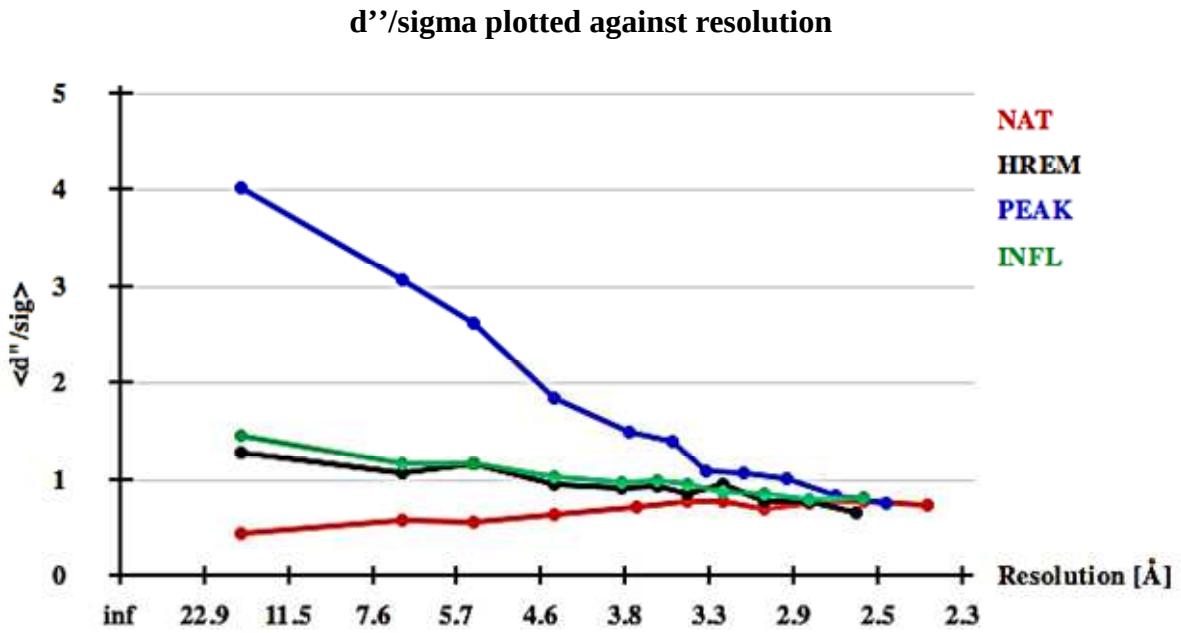


Figure 8.7 Results from SHELX C showing anomalous signal from the four BPSS0212 datasets. Graph of d''/sigma plotted against resolution, a value of 1.2 or above indicates good anomalous signal.

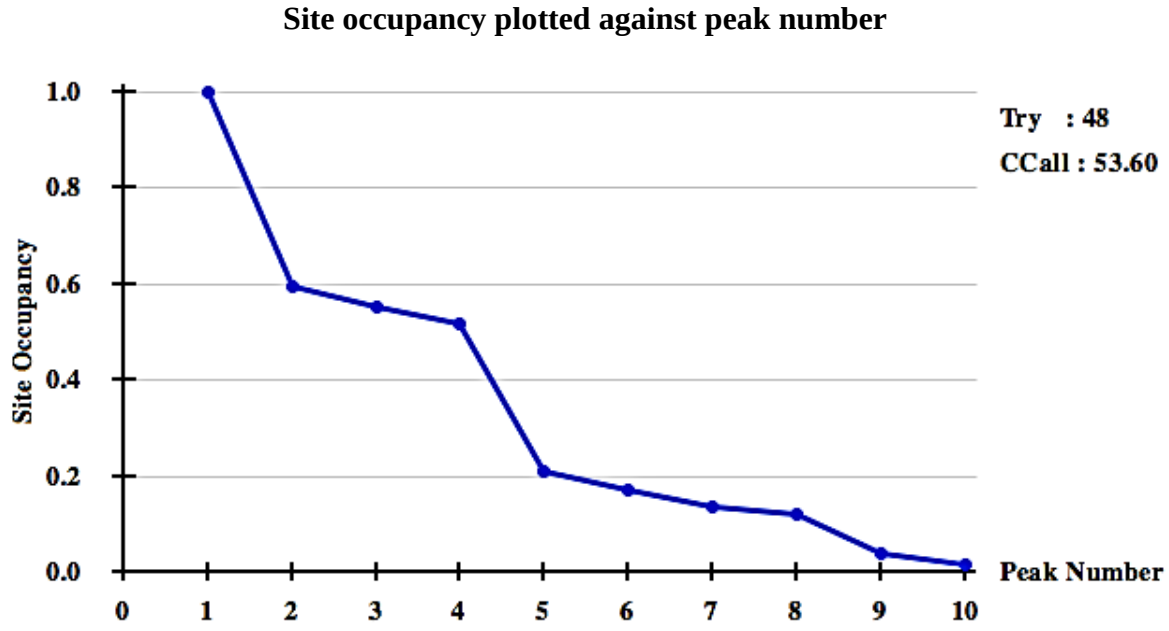
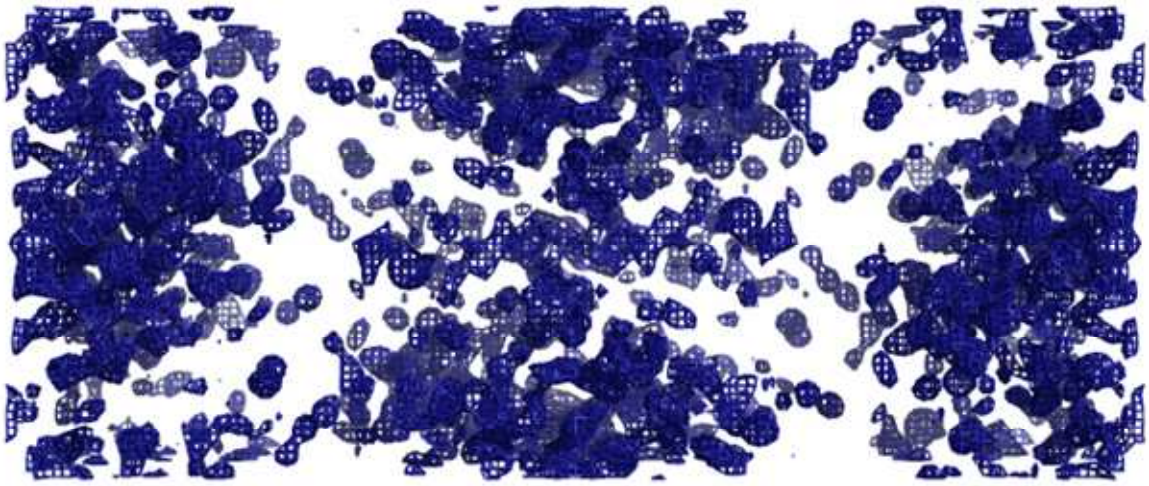


Figure 8.8 Results from SHELX D for BPSS0212 MAD experiment showing the best solution. Peaks are predicted to have occupancies of 0.99, 0.59, 0.54 and 0.51 before a drop to 0.20 and below.

(a) Original substructure hand



(b) Inverted substructure hand

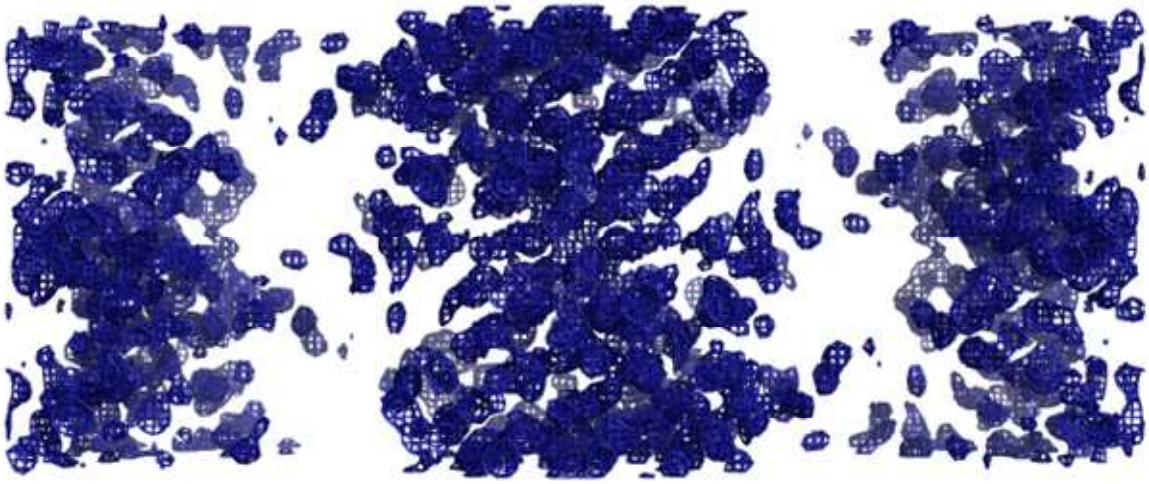


Figure 8.9 Sample region of electron density for the original hand and inverted hand solutions for the BPSS0212 selenium MAD phasing experiment. Both maps are contoured at 1.5 sigma and show the same cross section of density covering the entire unit cell. **a** The initial electron density map output from SHELX E for the original substructure hand. **b** The initial electron density map output from SHELX E for the inverted substructure hand.

9.0 Studies on the protein BPSS0213

This section covers the purification, crystallisation and preliminary data collection for the target BPSS0213.

9.1 Purification of BPSS0213

Approximately 3 g of cell paste was resuspended in 30 ml 50 mM TRIS pH 9.0 and disrupted by sonication. Cell debris and insoluble protein was removed by centrifugation at 70,000 g for 15 minutes and the supernatant was loaded onto a DEAE-Sepharose fast flow column equilabrated with 50 mM TRIS pH 9.0. A 200 ml gradient from 0 to 500 mM NaCl was applied to the column and 8 ml fractions were collected (figure 9.1 a). Fractions were analysed by SDS-PAGE (figure 9.2) and the peak fractions containing BPSS0213 were concentrated to 1.5 ml using a Vivaspin concentrator with a 10 kDa MWCO. This was then loaded onto a 1.6 x 60 cm Superose 6 gel filtration column equilibrated in 50 mM TRIS pH 8.0 and 500 mM NaCl and eluted using the same buffer collecting 2 ml fractions (figure 9.1 b). Fractions were analysed by SDS-PAGE (figure 9.2) and those containing BPSS0213 were combined. The buffer was exchanged for 100 mM NaCl and 10 mM TRIS pH 8.0 and the protein was concentrated to 6 mg ml^{-1} for use in crystallisation trials using a Vivaspin concentrator with a 10 kDa MWCO.

9.1.1 Protein sample analysis for BPSS0213

The elution profile for BPSS0213 from the gel filtration column shows a single peak roughly corresponding to a dimeric form of the protein (figure 9.1 b). SDS-PAGE analysis of the purification shows the sample for crystallisation studies was over 90 % pure but suffering from degradation even during the timescale of the purification process (figure 9.2).

9.2 Crystallisation of BPSS0213

Four initial 96 condition robot screens, the JCSG+, PACT, Pegs and Classics suites, were conducted using purified BPSS0213 at 6 mg ml^{-1} in 10 mM TRIS pH 8.0 and 100 mM NaCl. The screens were conducted using 200 nl of protein mixed with 200 nl of well solution and the trays were incubated at 17 ºC. Two hits were found in the JCSG+ robot crystal screen after four months. Tiny crystals were found in condition JCSG D6, 200 mM Magnesium chloride, 100 mM TRIS pH 8.5 and 20 % (w/v) PEG 8,000. A single crystal was found in JCSG D7, 200 mM Lithium sulphate, 100 mM TRIS pH 8.5 and 40 % (v/v) PEG 400. Optimisation

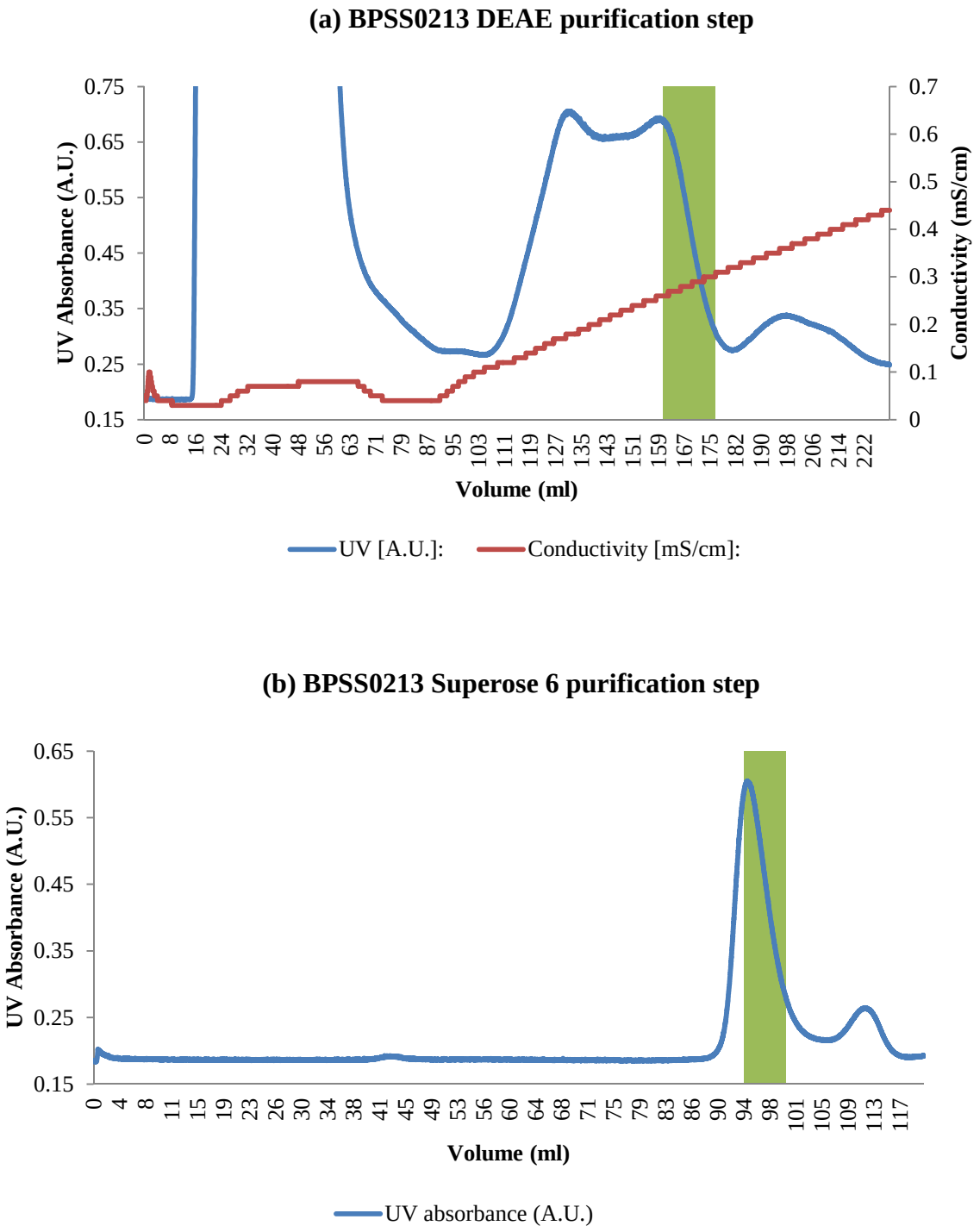


Figure 9.1 Chromatogram traces for the purification of BPSS0213. **a** DEAE purification step showing column loading and elution. 8 ml fractions were collected starting at the beginning of the gradient at 40 ml. **b** Gel filtration purification step showing elution with 2 ml fractions collected throughout. For both traces, green highlighted regions indicate volumes taken for subsequent purification steps or as pure protein.

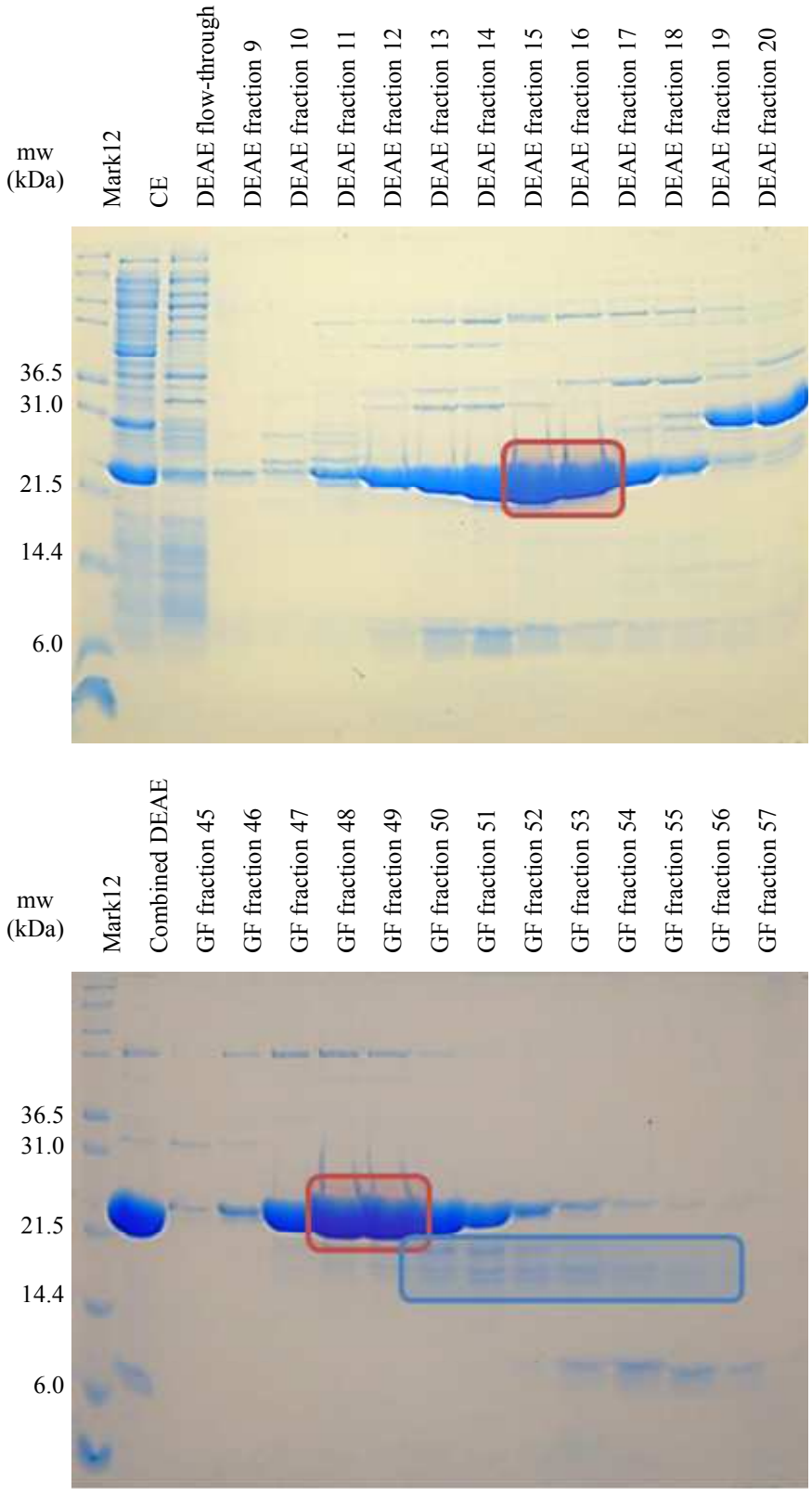
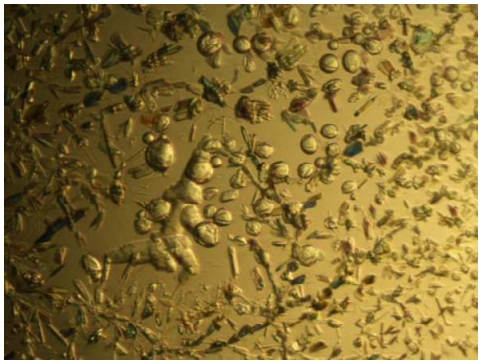


Figure 9.2 SDS-PAGE gels showing the purification of BPSS0213. The molecular weight of BPSS0213 is 21.7 kDa and the highlighted bands indicate the protein in fractions taken for subsequent purification steps or as pure protein. The blue highlighted bands indicate degradation products of BPSS0213.

trials were conducted for JCSG D6 by altering the PEG concentration (10 – 30 % (w/v)), protein concentration (3 – 10 mg ml^{-1}) and pH (8.0 – 9.0) but resulted in no further crystals growing. A further robot screen around this initial condition using the 96 well Optisalts screen also resulted in no crystals forming. Attempts to optimise JCSG D7 by altering the PEG concentration (25 – 55 % (w/v)), protein concentration (3 – 10 mg ml^{-1}) and pH (8.0 – 9.5) using 2 µl protein solution and 2 µl well solution hanging drop trials produced drops containing crystals alongside possible quasi crystals and fungal growth (figure 9.3) while the initial hit later turned out to be a salt crystal.



JCSG D7 – Hanging drop
200 mM Lithium sulphate
100 mM TRIS pH 8.5
40 % (w/v) PEG 400
Drop size 2 µl protein + 2 µl well solution

Figure 9.3 Photograph of BPSS0213 crystals.

9.3 Data collection for BPSS0213

A number of crystals were looped by dragging a tennis racket loop through the crystallisation drop before being flash frozen in a stream of gaseous nitrogen at 100 ºK and mounted onto the detector. Diffraction analysis was conducted using an in-house Rigaku MM007 rotating anode generator and a MAR 345 research image plate. A single image was taken using an exposure time of 900 s and a crystal to detector distance of 200 mm (figure 9.4). The image contained some low resolution reflections not associated with ice rings suggesting that there were ordered protein crystals present in the crystallisation drop. Attempts to optimise the crystallisation conditions have not yet been conducted.

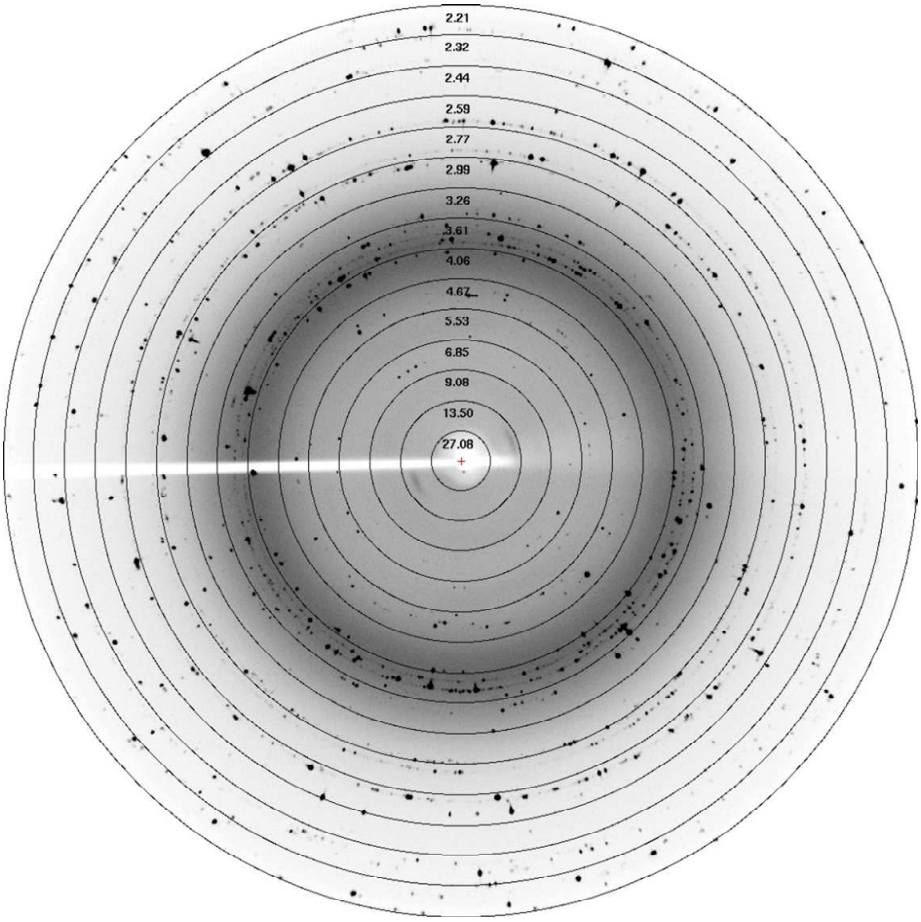


Figure 9.4 Diffraction image for a number of crystals of BPSS0213. Ice rings are found at 3.897, 3.669, 3.441, 2.671 and 2.249 Å, the presence of diffraction at lower resolutions indicate the presence of protein crystals.

Chapter three

Introduction to, and results from, an on-going project to elucidate the structure and mechanism of proteins in the thioredoxin system from *Burkholderia pseudomallei*

Section 10 The thioredoxin system is essential and represents a potential drug target for bacterial diseases

Section 11 Studies on the protein BPSL1497

10.0 The thioredoxin system is essential and represents a potential drug target for bacterial diseases

Intracellular proteins exist in a very reducing environment and cysteine residues are maintained in the sulphydryl (-SH) form with disulphide bridges (-S-S-) existing very rarely. There are two systems responsible for the maintenance of protein thiol states, the glutathione and thioredoxin systems. The thioredoxin system in *Burkholderia pseudomallei* comprises two proteins, thioredoxin (BPSL1497) and thioredoxin reductase (BPSL2605).

10.1 Thioredoxin

Thioredoxins are a highly conserved class of small (10 – 12 kDa) peptides, present in all organisms. They all share the same characteristic tertiary structure containing a dithiol:disulphide active site with the motif CXXC (usually CGPC). Thioredoxin has several essential roles within the prokaryotic cell alongside its role as a protein disulphide oxidoreductase [167]. Thioredoxin also works as a high capacity electron donor to a number of important redox enzymes including ribonucleotide reductase and thioredoxin peroxidase. The protein has a further role in controlling the transcription of some genes by reducing key cysteine residues in a number of transcription factors. A BLAST search [165] for the thioredoxin protein sequence from *Burkholderia pseudomallei* demonstrates that it is clearly homologous to the protein from *Escherichia coli* sharing 65 % of residue identities (figure 10.1 a).

10.2 Thioredoxin reductase

Thioredoxin reductase represents the second component of the thioredoxin system. It is a flavoprotein responsible for the NADPH dependent reduction of thioredoxin to regenerate its biologically active disulphide bond. Two forms of thioredoxin reductase are known and have evolved separately: prokaryotes, archea and the majority of plants have a low molecular weight (approximately 35 kDa) thioredoxin reductase while higher eukaryotes have a high molecular weight (approximately 55 kDa) variety that contains a seleno-cysteine residue and operates by a different mechanism. The difference between the two forms of thioredoxin reductase, and the essential nature of the system, marks the protein as a potential drug target in fighting bacterial infections. The mechanism of low molecular weight thioredoxin reductase has been extensively studied [168-176]. The thioredoxin reductase transfers reducing equivalents from NADPH to its bound flavin, which reduces an internal disulphide bond, which can then reduce the disulphide bond of thioredoxin (figure 10.2). A BLAST

search [165] for the thioredoxin reductase protein sequence from *Burkholderia pseudomallei* demonstrates that it is clearly homologous to the protein from *Escherichia coli* sharing 68 % of residue identities (figure 10.1 b).

(a) Thioredoxin (BPSL1497)

Score = 154 bits (390), Expect = 6e-49, Method: Compositional matrix adjust.
Identities = 69/107 (64%), Positives = 89/107 (83%), Gaps = 0/107 (0%)

```
BP      2  SEQIKHISDASFEQDVVKSDKPVLLDFWAEWCGPCKMIAPILDEVAKDYGDKLQIAKINV 61
           S++I H++D SF+ DV+K+D  +L+DFWAEWCGPCKMIAPILDE+A +Y  KL +AK+N+
EC      1  SDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNI 60

BP     62  DENQATPAKFGVRGIPTLILFKNGAVAAQKVGALSKSQLTAFLDSHL 108
           D+N  T  K+G+RGIPTL+LFKNG VAA KVGALSK QL  FLD++L
EC     61  DQNPGTAPKYGIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDANL 107
```

(b) Thioredoxin reductase (BPSL2605)

Score = 447 bits (1150), Expect = 2e-157, Method: Compositional matrix adjust.
Identities = 215/319 (67%), Positives = 257/319 (81%), Gaps = 1/319 (0%)

```
BP      3  TPKHAKVLILGSGPAGYTAAVYAARANLSPLLITGIAQGGQLMTTTDVENWPADADGVQG 62
           T KH+K+LILGSGPAGYTAAVYAARANL P+LITG+ +GGQL TTT+VENWP D + + G
EC      2  TTKHSKLLILGSGPAGYTAAVYAARANLQPVLITGMEKGGQLTTTTEVENWPGDPNDLTG 61

BP     63  PELMQRFLAHAQRFNTEIVFDHIHTAKLHEKPIRLIGDSGEYTCDSLIIATGASAQYLGL 122
           P LM+R   HA +F TEI+FDHI+   L  +P RL GD+GEYTCD+LIIATGASA+YLGL
EC     62  PLLMERMHEHATKFETEIIFDHINKVDLQNRPFRLNGDNGEYTCDALIIATGASARYLGL 121

BP    123  QSEEAFMGRGVSACATCDGFFYRGQNVAVVGGGNTAVEEALYLTGIAKKVTVIHRRDKFR 182
            SEEAF GRGVSACATCDGFFYR Q VAV+GGGNTAVEEALYL+ IA +V +IHRRD FR
EC    122  PSEEAFKGRGVSACATCDGFFYRNQKVAVIGGGNTAVEEALYLSNIASEVHLIHRRDGFR 181

BP    183  AEPILVDRLLEKEKEGAVEIKWDHVLDEVTGDDSGVSGVRIKHV-TTGATEDVAVQGLFI 241
           AE IL+ RL++K + G + +  +  L+EVTGD  GV+GVR++    +   E + V GLF+
EC    182  AEKILIKRLMDKVENGNIILHTNRTLEEVTGDQMGVTGVRLRDTQNSDNIESLDVAGLFV 241

BP    242  AIGHKPNTDIFKGQLEMKDGYIITNSGLSGNATGTSVPGVFAAGDVQDHIYRQAITSAGT 301
           AIGH PNT IF+GQLE+++GYI   SG+ GNAT TS+PGVFAAGDV DHIYRQAITSAGT
EC    242  AIGHSPNTAIFEGQLELENGYIKVQSGIHGNATQTSIPGVFAAGDVMDHIYRQAITSAGT 301

BP    302  GCMAALDAQRYLESLHDHK 320
           GCMAALDA+RYL+ L D K
EC    302  GCMAALDAERYLDGLADAK 320
```

Figure 10.1 BLAST search results for BPSL1497 and BPSL2605. The two sequences designated BP are from *Burkholderia pseudomallei* and those designated EC are from *Escherichia coli*.

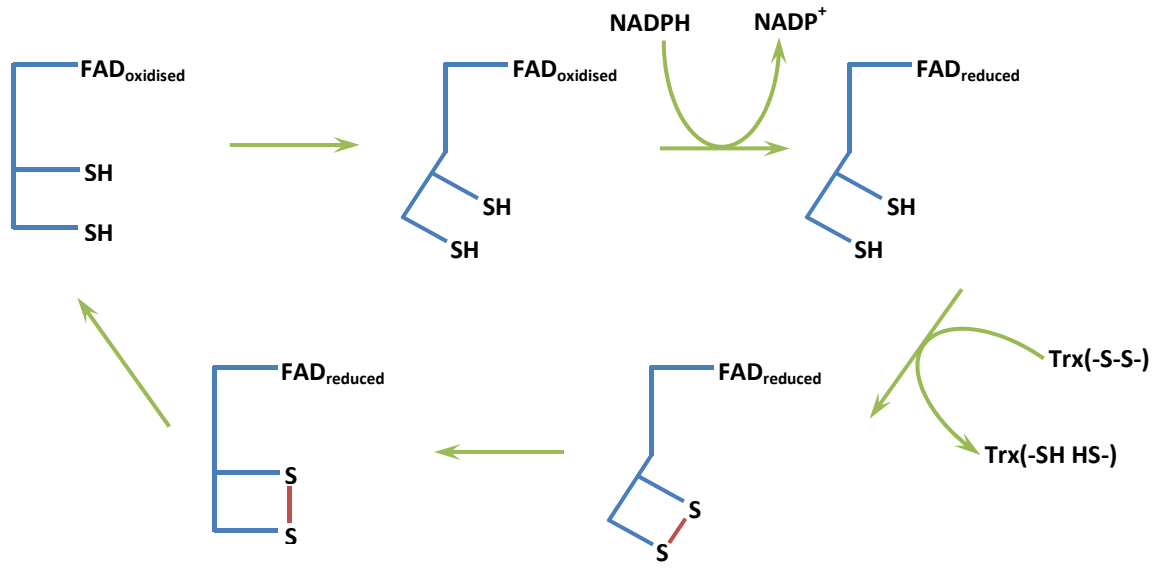


Figure 10.2 Schematic representation of the catalytic cycle of thioredoxin reductase. There are two different conformations of the enzyme represented by the straight and bent forms. The cycle can be thought of to begin in the FO conformation with the enzyme in a 2 electron reduced state with an oxidised flavin and reduced cysteine residues. The enzyme then alters conformation to the FR state, allowing interaction with NADPH and the reduction of the flavin moeity to produce a four electron reduced form of the enzyme. In this conformation the enzyme is also able to interact with thioredoxin (Trx), replacing the disulphide in Trx with two sulphdryl groups by oxidising its own cysteine residues to form a disulphide. The enzyme then rearranges into its original conformation allowing the reduction of the disulphide by the flavin. Figure adapted from "Crystal structure of reduced thioredoxin reductase from Escherichia coli: Structural flexibility in the isoalloxazine ring of the flavin adenine dinucleotide cofactor" [173]

10.3 Thioredoxin reductase has two distinct conformations

As part of the mechanism employed by this enzyme, a large conformational change has to occur. The two conformers of the protein are referred to as the flavin oxidising (FO) and flavin reducing (FR) forms.

10.3.1 The FO conformation of thioredoxin reductase

The first structure solved for a low molecular weight thioredoxin reductase was from *Escherichia coli* [178]. The structure contained a dimer of two identical subunits, each consisting of two distinct domains. One domain contains the active site cysteine residues and a NADPH binding site and is therefore termed the NADPH binding domain. The other domain, designated the FAD binding domain, contains the non-covalently bound flavin moiety. In the structure the isoalloxazine ring of the FAD is closely associated with the active

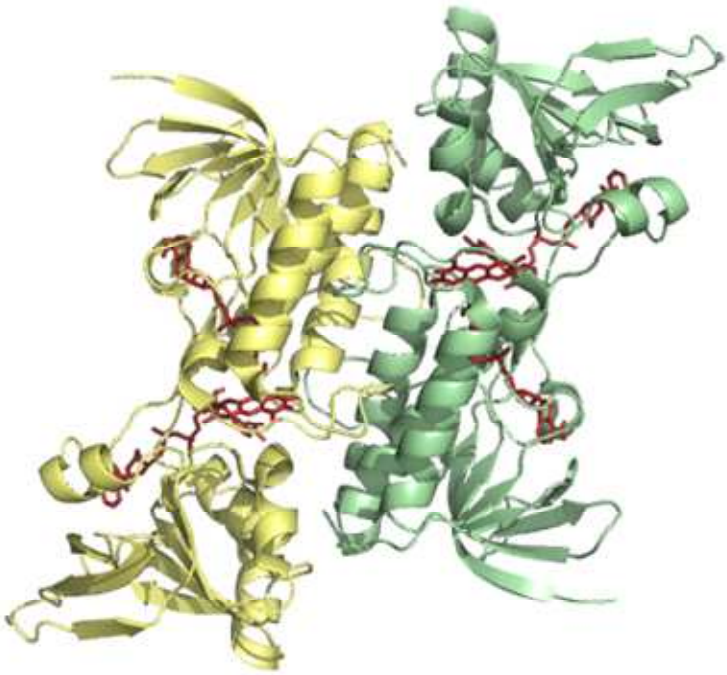
site cysteines, allowing the transfer of reducing equivalents from the FAD to the active site disulphide. However the NADPH binding site is distant from the flavin and the active site cysteine residues are buried and therefore inaccessible to the enzymes thioredoxin substrate. It was proposed that the protein must undergo a conformational change as part of its mechanism to allow the transfer of reducing equivalents along the chain from NADPH, to FAD, to the internal disulphide and finally to thioredoxin. As the solved conformation allowed the transfer of reducing equivalents from the FAD to the active site disulphide it was classified as the flavin oxidising (FO) form of the protein.

10.3.2 The FR conformation of thioredoxin reductase

The two domains of the enzyme share structural homology to the corresponding domains of other members of the pyridine nucleotide disulphide oxidoreductase flavoprotein family, such as glutathione reductase. However in glutathione reductase the organisation of the domains is drastically different, with an extra interface domain involved in dimerisation. The active site cysteine residues are also located on different domains, and a conformational change is not required as part of the enzymes mechanism [179]. A model of an alternative, flavin reducing (FR), conformation of thioredoxin reductase was proposed based on the structure of glutathione reductase [178]. In the model the two domains were rotated with respect to each other, allowing the FAD and NADPH molecules to interact and exposing the active site cysteine residues to the solvent. The proposed domain rotation appeared to be plausible as the two domains made relatively few contacts with other regions of the dimeric enzyme and that no barriers to the change existed that couldn't be resolved by the adjustment of sidechains [178]. The requirement for a conformational change was supported by other studies involving measuring the kinetic, spectroscopic and redox properties of various mutated forms of the enzyme, cross-linked domains preventing conformational switching and complexes created by cross-linking thioredoxin reductase with thioredoxin [180-184]. A 3.0 Å structure relating to the FR conformation has since been solved for the protein from *Escherichia coli* [185]. Mutant variants of thioredoxin and thioredoxin reductase with one remaining active site cysteine each were used to form a stable covalent complex. The disulphide cross-link prevents the two proteins dissociating with thioredoxin acting as a "doorstop", locking thioredoxin reductase into the FR conformation [181]. The structure supported the predicted model with the two domains rotated 67º with respect to each other. The active site cysteine residues are clearly available for interaction with thioredoxin and the structure was solved in the presence of $AADP^+$, which is found in the NADPH binding site with its pyridine ring

stacked against the isoalloxazine ring of the FAD. The FR conformation is therefore able to pass reducing equivalents between NADPH and FAD, and the active site cysteine residues and its substrate.

(a) Thioredoxin reductase in the FO conformation



(b) Thioredoxin reductase locked in the FR conformation by thioredoxin

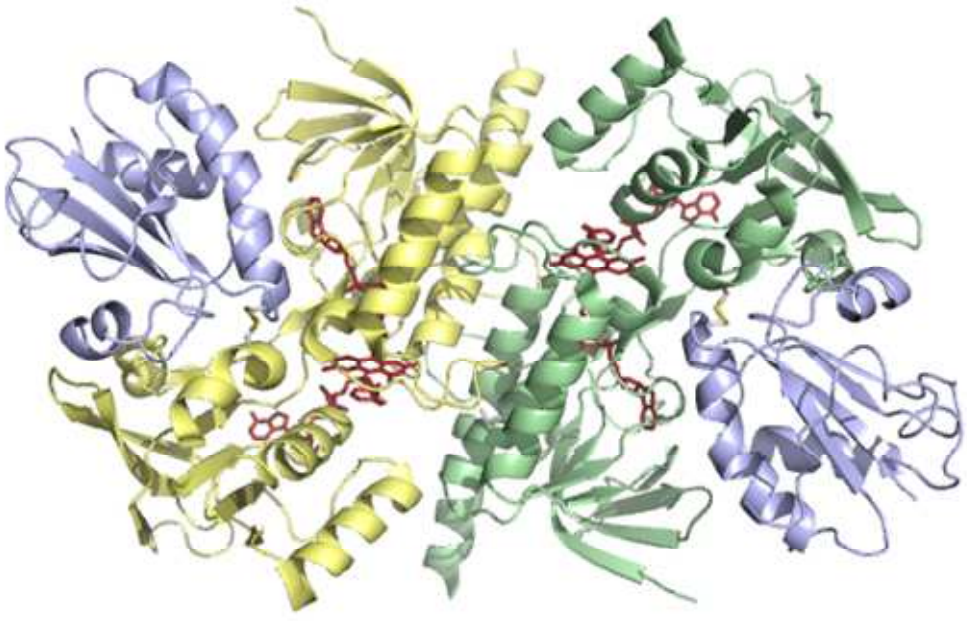


Figure 10.3 The two conformations of thioredoxin reductase from *Escherichia coli*.

10.4 Project aim

A crystal structure of the oxidized FO conformation of thioredoxin reductase from *Burkholderia pseudomallei* had previously been determined to 1.45 Å within the Sheffield crystallography group (Martin Bush, Thesis 2009). The aim of this project was to solve a high resolution structure of thioredoxin from *Burkholderia pseudomallei* to lead into work determining the FR conformation of thioredoxin reductase at high resolution in the future. A high resolution of this complex, representing an intermediate of the catalytic mechanism, would resolve some of the remaining ambiguities in the proposed mechanism of thioredoxin reductase. This work will lead to a better understanding of the system in *Burkholderia pseudomallei* and, due to the high levels of conservation, across all prokaryotic organisms. This information could then be used to inform the design of inhibitors to thioredoxin reductase providing novel antibacterial compounds.

11.0 Studies on thioredoxin from *Burkholderia pseudomallei*

This section describes the cloning, overexpression, purification, crystallisation, data collection, processing and resulting structure for thioredoxin from *Burkholderia pseudomallei* (BPSL1497) solved by molecular replacement.

11.1 Cloning of BPSL1497

BPSL1497 was amplified from *Burkholderia pseudomallei* strain D286 genomic DNA by PCR using BioMix Red (Bioline) and specific primers (Eurofins) (table 11.1). PCR products were purified using agarose gel electrophoresis and a QIAquick® Gel Extraction Kit (Qiagen) by standard protocols. The purified PCR product was ligated into a pETBlue-1 vector using an AccepTor vector kit (Novagen) and used to transform Novablue competent cells (Novagen) which were plated on LB-agar with the addition of 100 µg ml^{-1} carbenicillin 70 µg ml^{-1} X-gal and 80 µM IPTG for selection and blue-white screening of colonies. Gene presence and orientation was confirmed by colony PCR on white colonies (figure 11.1) before plasmids were recovered by miniprep and sequenced (Geneservice) to ensure the gene sequence was correct.

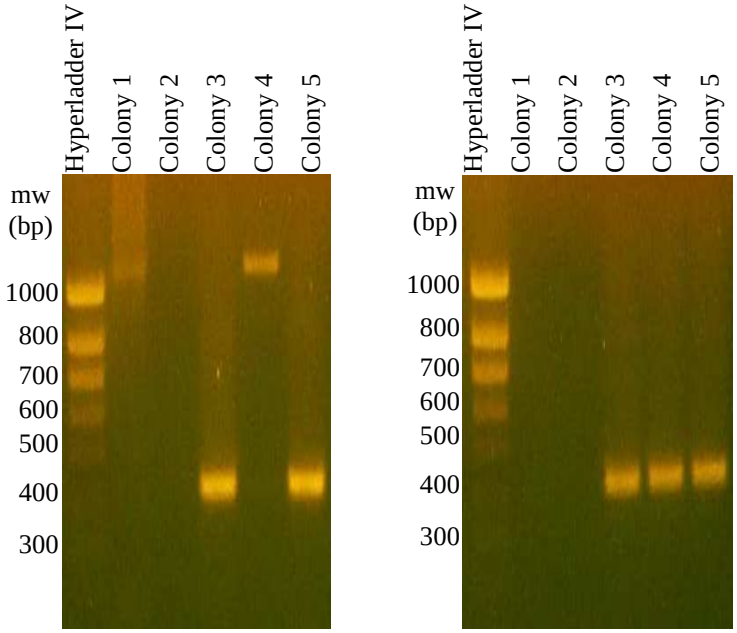


Figure 11.1 Agarose gels showing the results of colony PCR for BPSL1497 cloning using different primers. Bpsl1497F and Bpsl1497R correspond to the start and end of the gene, pETBlueUP and pETBlueDOWN correspond to regions on the plasmid. The expected band size using pETBlueUP and Bpsl1497R was 422 base pairs while Bpsl1497F and pETBlueDOWN should produce a band at 405 base pairs if the gene has inserted correctly. Colonies 3 and 5 produce the correct sized bands for both reactions.

Oligoname	Sequence
bpsl1497 F	5’-ATGAGCGAACAGATCAAACACATCA-3’
bpsl1497 R	5’-GCATCCGGTTTTAGAGAGATGGCTGT-3’

Table 11.1 Primers used for the PCR amplification of BPSL1497 from genomic DNA.

11.2 Protein overexpression and purification for BPSL1497

11.2.1 Overexpression

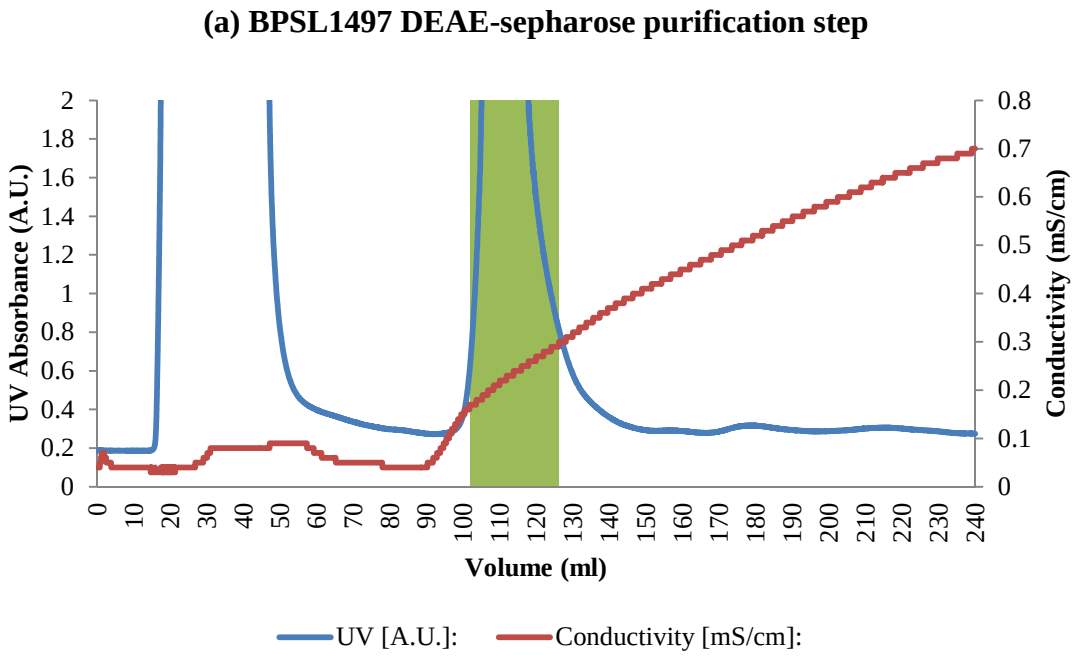
Plasmids containing correctly sequenced genes were used to transform Tuner (DE3) pLacI competent cells (Novagen) for overexpression. A small-scale over expression trial was performed for BPSL1497 to find post induction conditions that would produce soluble protein. The best post induction condition was found to be 1 mM IPTG at 37 ºC for 20 hours.

11.2.2 Purification

Approximately 4 g of cell paste was resuspended in 40 ml 50 mM TRIS pH 9.0 and disrupted by sonication. Cell debris and insoluble protein was removed by centrifugation at 70,000 g for 15 minutes and the supernatant loaded onto a DEAE-Sepharose fast flow column equilibrated with 50 mM TRIS pH 9.0. A 200 ml gradient from 0 to 300 mM NaCl was then applied to the column and 8 ml fractions were collected (figure 11.2a). Fractions were analysed by SDS-PAGE (figure 11.3) and fractions containing BPSL1497 were pooled and concentrated to 1.5 ml using a Vivaspin concentrator with a 5 kDa MWCO. This was then loaded onto a 1.6 x 60 cm Superose 6 gel filtration column equilibrated in 50 mM TRIS pH 8.0 and 500 mM NaCl and eulted using the same buffer collecting 2 ml fractions (figure 11.2b). Peak fractions containing BPSL1497 were combined. The buffer was exchanged for 10 mM TRIS pH 8.0 using a Vivaspin concentrator with a 5 kDa MWCO and the protein was concentrated to 10 mg ml^{-1} for use in crystallisation trials using the same Vivaspin concentrator. The overall yield of protein was 12 mg and was estimated by SDS-PAGE to be over 95 % pure (figure 11.3).

11.2.3 Purification analysis

The elution profile for BPSL1497 from the gel filtration column shows a single large peak roughly corresponding to a monomeric form of the protein (figure 12.2b).



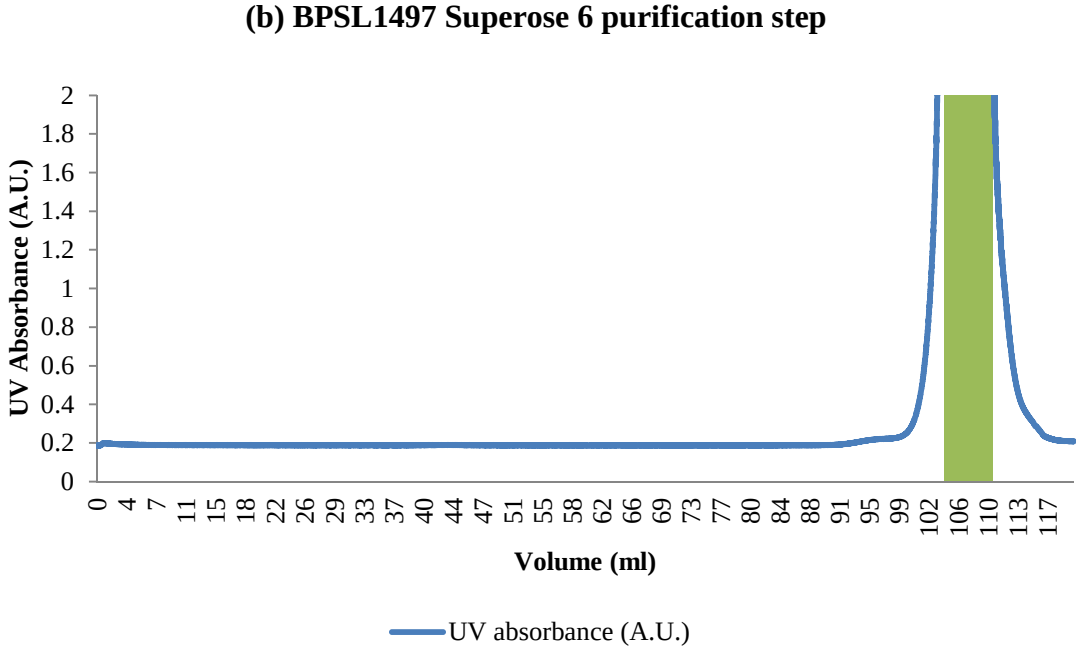


Figure 11.2 Chromatogram traces for the purification of BPSL1497. **a** DEAE purification step showing column loading and elution. 8 ml fractions were collected starting at the beginning of the gradient at 32 ml. **b** Gel filtration purification step showing elution with 2 ml fractions collected after the void volume of 44 ml. For all traces, green highlighted regions indicate volumes taken for subsequent purification steps or as pure protein.

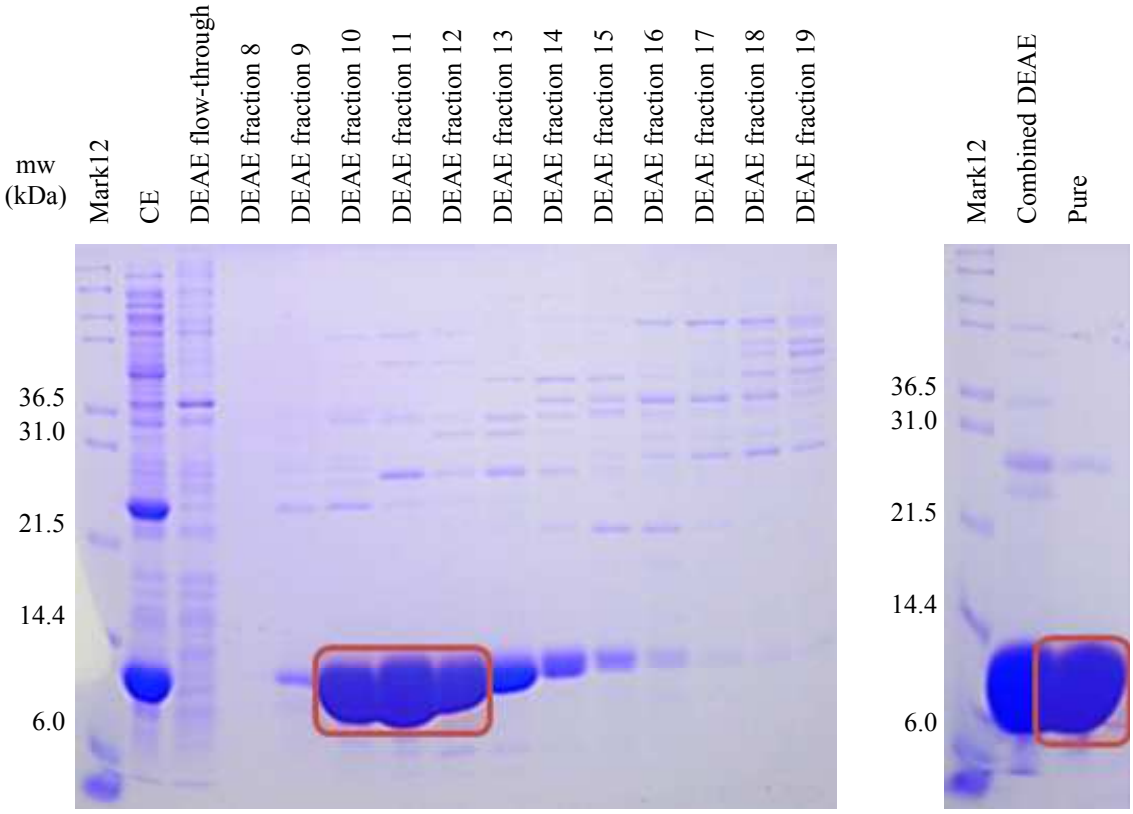
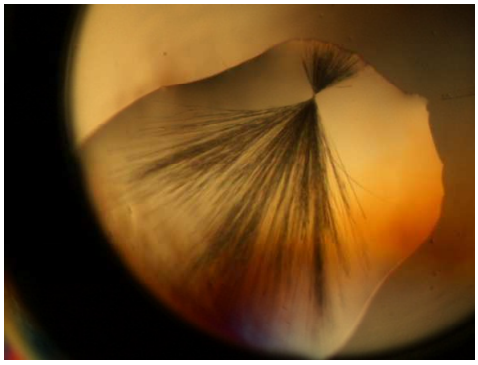


Figure 11.3 SDS-PAGE gel showing the purification of BPSL1497. The molecular weight of BPSL1497 is 11.8 kDa and the highlighted bands indicate the protein in fractions taken for subsequent purification steps or as pure protein.

11.3 Protein crystallisation for BPSL1497

Two initial 96 condition robot screens, the JCSG+ and PACT suites, were conducted using purified BPSL1497 at 10 mg ml^{-1}. 200 nl of protein was mixed with 200 nl of well solution using a matrix hydra II plus one (Thermo Scientific) robot and the trays were incubated at 17 ºC. Several hits were found in the screens with varying morphologies from needle clusters to well defined cubic crystals in JCSG+ condition C7, 100 mM sodium acetate pH 4.5, 200 mM zinc acetate and 10 % (w/v) PEG 3,000 (figure 11.4). Attempts to optimise the crystals by altering the PEG concentration (5 – 20 % (w/v)) using 5 µl protein solution and 5 µl well solution found optimum conditions were 100 mM sodium acetate pH 4.5, 200 mM zinc acetate and 7 % (w/v) PEG 3,000. This condition produced several large crystals for data collection.



PACT A10– Robot screen

200 mM Magnesium chloride

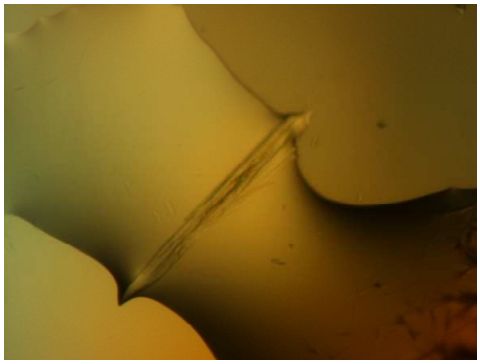
100 mM Sodium acetate pH 5.0

20 % (w/v) PEG 6000

PACT C1 – Robot screen

100 mM PCB buffer pH 4.0

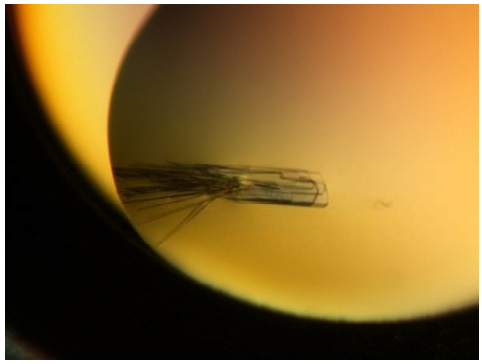
25 % (w/v) PEG 1500



PACT C2 – Robot screen

100 mM PCB buffer pH 5.0

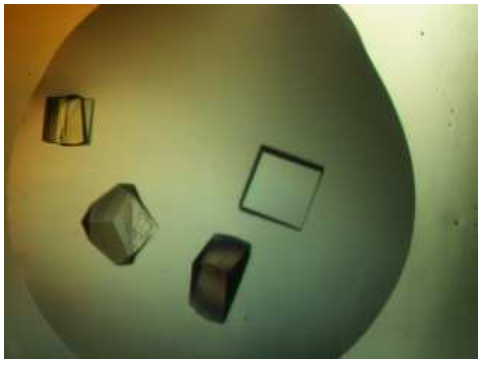
25 % (w/v) PEG 1500



JCSG+ C6 – Robot screen

100 mM Phosphate-citrate pH 4.2

40 % (w/v) PEG 300



JCSG+ C7 – Robot screen

200 mM Zinc acetate

100 mM Sodium acetate pH 4.5

10 % (w/v) PEG 3000

Figure 11.4 Photographs of BPSL1497 crystals.

11.4 BPSL1497 initial data

11.4.1 Initial data collection

Crystals were selected for data collection based on their size and quality from the optimisation trials. Crystals were looped and placed into a variety of cryoprotectant solutions made using the well solution with the addition of differing concentrations of ethylene glycol (15 – 30 % (v/v)), PEG 400, (10 – 30 % (v/v)), glycerol (10 – 30 % (v/v)) or MPD (15 – 30 % (v/v)). Once crystals were placed into any of the cryoprotectant solutions they began to show signs of distress with the crystals visibly cracking. They were then looped again and flash frozen in a stream of gaseous nitrogen at 100 ºK and mounted onto the detector. Initial diffraction analysis was conducted in order to determine the diffraction quality of several crystals using an in-house Rigaku MM007 rotating anode generator and a MAR 345 research image plate. Two images were collected 90º apart with 1º oscillation and a crystal to detector distance of 250 mm for a variety of crystals and cryoprotectants. The diffraction from these crystals was very poor with obvious multiple lattices and smeared diffraction spots. A new set of crystals were looped and flash frozen without the addition of any cryoprotectant. The diffraction from these crystals was of exceptional quality and due to the strength of diffraction the presence of ice rings was not a problem. A number of crystals that diffracted beyond 2 Å on the home source were saved and taken to a synchrotron at the I03 beamline of the Diamond light source, Oxford. Three initial test images were taken 45º apart with 1º oscillation to ensure crystal centering and to obtain a collection strategy using the auto-indexing and collection strategy components of Mosflm [146]. For the initial data set 900 images were collected with 0.1º oscillation per image using X-rays of 15000 eV at a crystal to detector distance of 351.7 mm using an ADSC Q315r detector. Data extending to 1.3 Å were collected (figure 11.5).

11.4.2 Initial data processing

The X-ray diffraction datasets were auto-processed at Diamond through the xia2 system using the 3dii mode [147]. Data were indexed and integrated by XDS and scaled by XSCALE [148]. The data indexed to space group $P4_12_12$ or $P4_32_12$, with the unit cell parameters a = b = 34.3 Å, c = 191.5 Å, (table 11.2). Although the dataset contained data to 1.3 Å, it was incomplete beyond 1.45 Å (figure 11.6). The asymmetric unit was predicted to contain one protein molecule with a Vm of 2.38 based on the probabilities of Matthews coefficients calculated using Mattprob [149] (table 11.3).

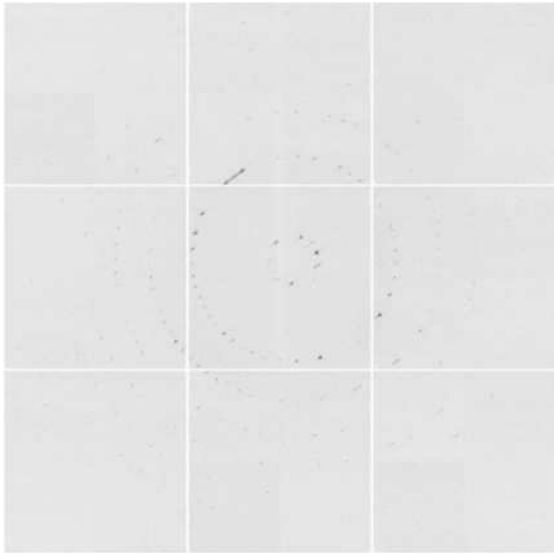


Figure 11.5 Diffraction image for BPSL1497 initial data collection.

Dataset	Initial crystal
Space group	$P4_32_12$
Unit cell parameters	
a (Å)	34.3
b (Å)	34.3
c (Å)	190.7
Energy (eV)	15000
Resolution range (Å)	34.31 – 1.28
Unique reflections	29794 (1691)
R_{merge}	0.054 (0.306)
R_{pim}	0.026 (0.249)
Completeness (%)	96.6 (77.5)
Multiplicity	5.7 (2.4)
Mean (I)/σ(I)	16.5 (2.4)

Table 11.2 Data collection statistics for initial protein crystal.

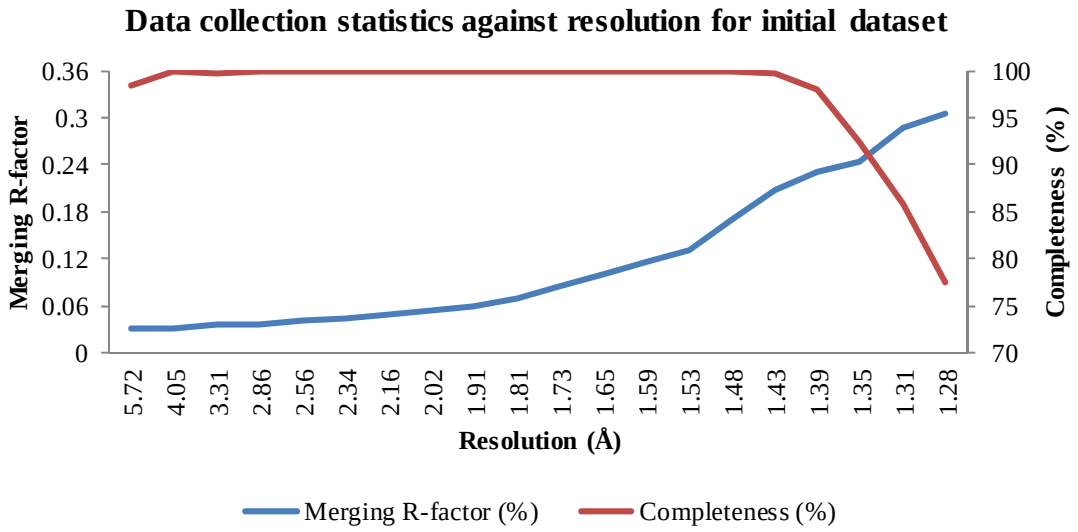


Figure 11.6 Data collection statistics against resolution for the BPSL1497 initial data. The graph shows R_{merge} and completeness for the dataset. The graph suggests the BPSL1497 initial data is incomplete but of good quality to the edge.

Molecules in the AU	Probability (based on resolution)	Probability (based on all proteins)	Vm ($Å^3$ / Da)	Solvent content (%)	Molecular weight (Da)
1	1.0000	1.0000	2.38	48.40	11815

Table 11.3 Matthews coefficient calculations and probabilities for BPSL1497. The results show there is one protein molecule inhabiting the asymmetric unit.

11.5 Phasing by molecular replacement

The Phyre 2 server [141] was used to create a suitable search model for molecular replacement using an *Escherichia coli* thioredoxin structure (PDB 3DXB) [186]. The search model was cut back to a poly alanine chain using chainsaw [150] before use for an automated search in Phaser [151]. The 1.5 Å dataset was input and Phaser was run in $P4_12_12$ and all the possible related space groups, searching for one molecule in the asymmetric unit. The best result was found using space group $P4_32_12$, the enantimorph of the original designation, which produced Z-scores of 7.1 and 15.1 for the rotation and translation functions respectively and an R-factor of 0.565 following refinement with REFMAC5 [152]. The resulting model and electron density were examined to confirm the result was unbiased by the search model with the presence of density relating to side chains visible (figure 11.10 a).

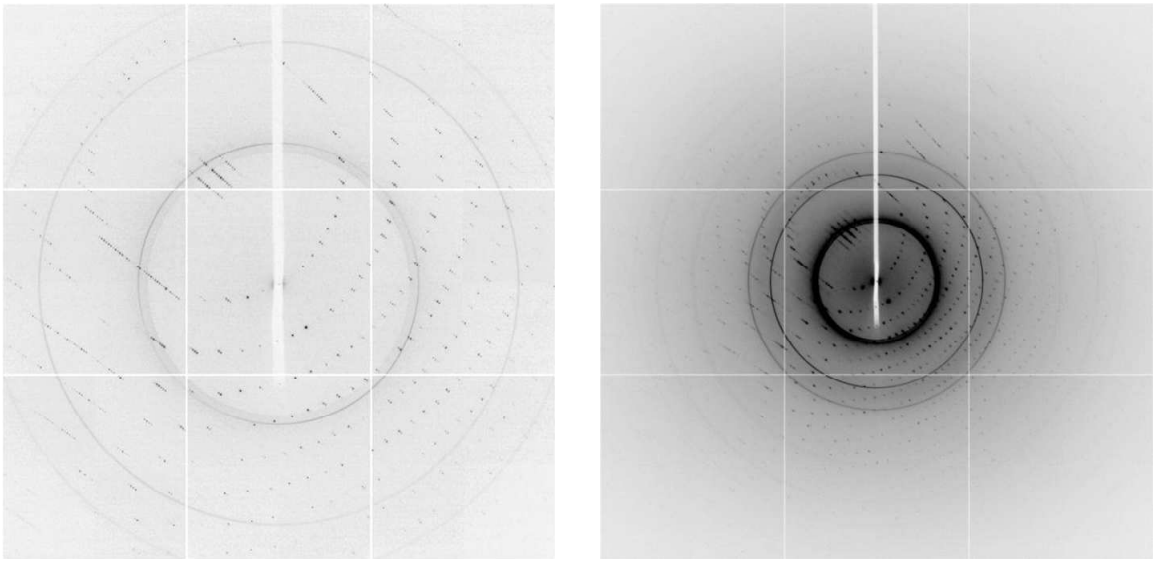
11.6 Model building and refinement

The starting model was then input into the Buccaneer pipeline along with the primary sequence of BPSL1497 to generate a model containing side chains. The pipeline consists of Buccaneer fast build [187] which produced a model with an R-factor of 0.514 (R_{free} 0.533) followed by 10 cycles of refinement using REFMAC5 [152] leading to a reduction in R-factor to 0.411 (R_{free} 0.449). Further refinement was conducted using iterative cycles of model building, refinement and evaluation using Coot [159] and REFMAC5 [152]. The process was repeated until a model for the initial data consisting of 105 residues, 201 water molecules and 3 zinc ions. The model had an R-factor of 0.1624 (R_{free} 0.1918) and agreed well with the electron density (figure 11.10 b).

11.7 High resolution data

11.7.1 High resolution data collection

The crystals of BPSL1497 were clearly capable of diffracting beyond the limits of the data collected. The presence of a long cell dimension, of approximately 192 Å along the c-axis, made collecting high resolution data difficult due to the proximity of the reflections to each other in the diffraction images. This is a particular problem for the low resolution data as the intensity of the reflections causes them to bleed into each other making assigning intensities for individual reflections problematic. In order to obtain higher resolution data, crystals that could diffract to high resolution with particularly low mosaicity were selected by analysis on the home source in Sheffield. These crystals were saved and taken to the I04 beamline of the Diamond light source, Oxford. Two initial test images were taken 90º apart with 1º oscillation to ensure crystal centering and to obtain a collection strategy using the auto-indexing and collection strategy components of Mosflm [146]. The beam used was slit down to the smallest possible size, 20 µm by 20 µm, to separate the reflections further. Two datasets were collected from the same crystal to best enable all data to be collected. The first set was collected using a low transmission and a relatively long detector distance of 289 mm in order to obtain a complete set of good quality low resolution data. The second dataset used a high transmission and relatively close detector distance of 127 mm, to maximise the measurable diffraction at high angles. For the two data sets 180 images were collected with 0.5º oscillation per image using X-rays of 12658 eV using an ADSC Q315r detector (figure 11.7).

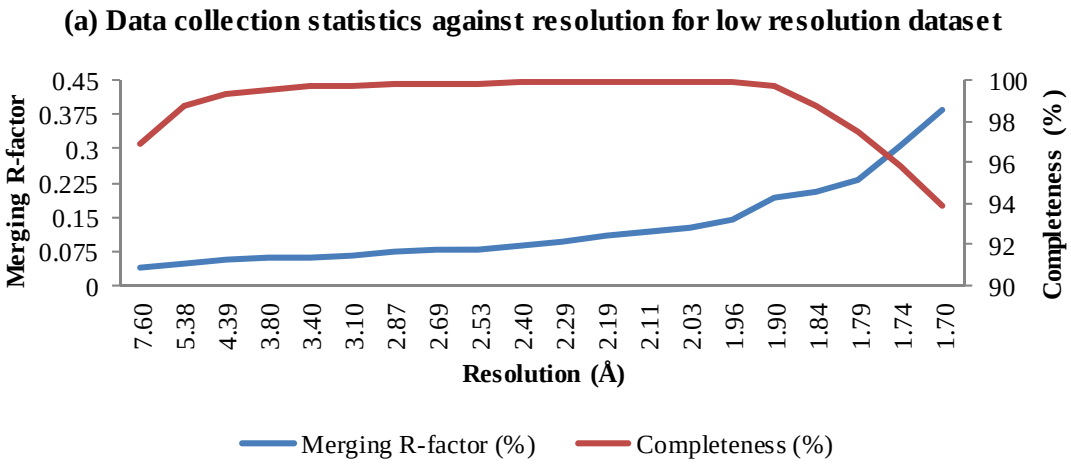


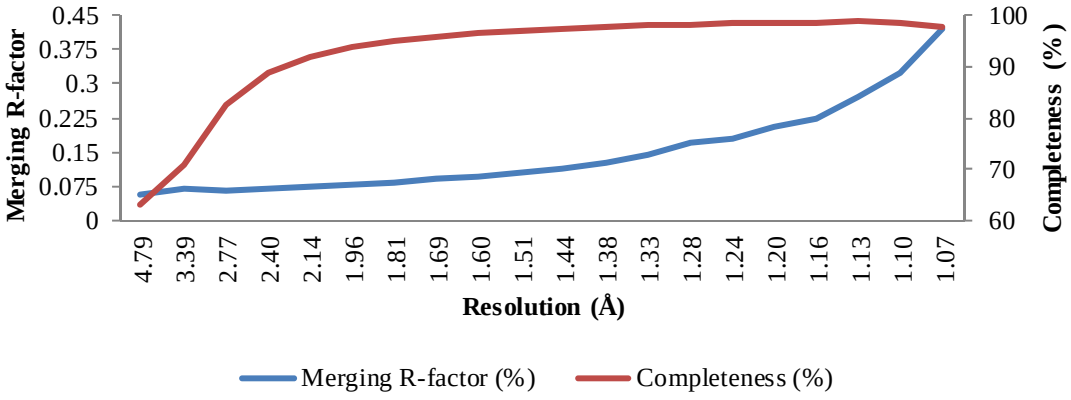
Low resolution dataset **High resolution dataset**

Figure 11.7 Diffraction images for BPSL1497 high resolution data collection.

Dataset	Low resolution	High resolution	Merged
Space group	$P4_32_12$	$P4_32_12$	$P4_32_12$
Unit cell parameters			
a (Å)	34.5	34.4	34.4
b (Å)	34.5	34.4	34.4
c (Å)	193.0	192.7	192.8
Energy (eV)	12658	12658	-
Resolution range (Å)	64.35 – 1.70	32.12 – 1.07	38.55 – 1.07
Unique reflections	13049 (705)	51425 (3413)	52072 (6933)
R_{merge}	0.077 (0.384)	0.089 (0.419)	0.110 (0.386)
R_{pim}	0.037 (0.256)	0.043 (0.280)	0.036 (0.227)
Completeness (%)	93.9 (70.3)	97.9 (90.4)	98.9 (92.7)
Multiplicity	5.5 (2.5)	5.9 (2.9)	7.2 (3.4)
Mean (I)/σ(I)	12.5 (2.1)	10.5 (2.3)	10.5 (2.5)

Table 11.4 Data collection statistics for higher resolution merged data.





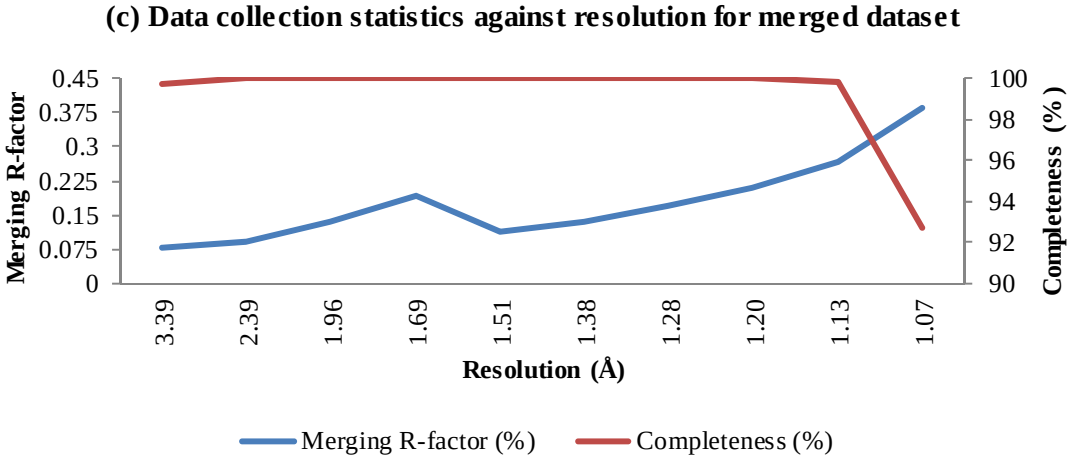


Figure 11.8 Data statistics against resolution against resolution for the BPSL1497 merged high resolution dataset and its constituent parts. The graphs show R_{merge} and completeness for each dataset. The low resolution data is complete to 1.95 Å, while the high resolution data is complete between 2.05 and 1.07 Å. The two datasets once merged and scaled together produce a dataset with good statistics to 1.10 Å.

11.7.2 High resolution data processing

The X-ray diffraction datasets were auto-processed at Diamond through the xia2 system using the 3dii mode [147]. Data were indexed and integrated by XDS and scaled by XSCALE [148]. Both datasets indexed to space group $P4_12_12$ or $P4_32_12$, with the unit cell parameters a = b = 34.5 Å, c = 193.0 Å, for the low resolution data and a = b = 34.4 Å, c = 192.7 Å, for the high resolution data (table 11.4). The two datasets were then merged and scaled together using sortmtz and scala [188] from the CCP4 suite of programmes [156]. Data from 38.57 to 2.00 Å from the low resolution dataset and 2.50 to 1.07 Å from the high resolution dataset were used and the resulting merged dataset had reasonable statistics (table 11.4) (figure 11.8).

11.8 Building the final model

The refined model from the 1.45 Å data was input into REFMAC5 [152] alongside the new 1.07 Å resolution merged data to allow phasing by direct refinement producing a model with an R-factor of 0.3333 (R_{free} 0.3446). Further refinement was conducted using iterative cycles of model building, refinement and evaluation using Coot [159] and REFMAC5 [152]. The final model (figure 11.9) consists of 105 residues, 4 zinc ions and 179 water molecules. The model had an R-factor of 0.187 (R_{free} 0.205) and agreed well with the electron density (figure 11.10 c).

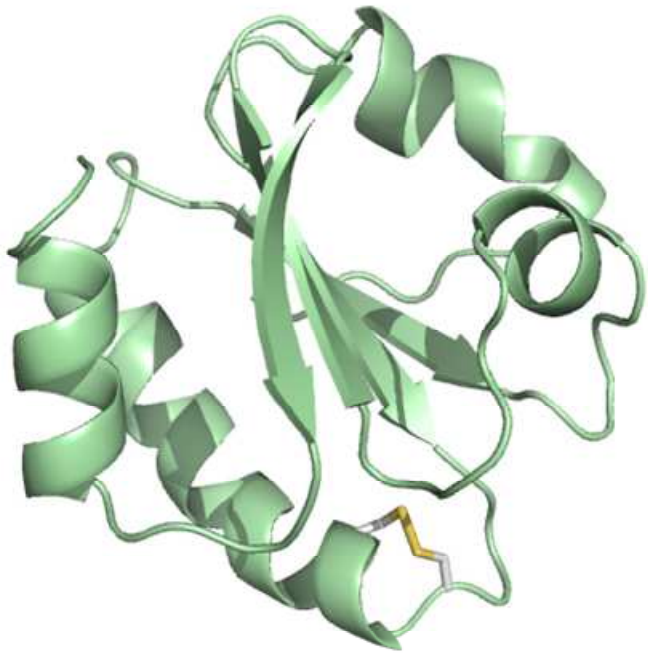


Figure 11.9 Cartoon representation of the overall fold of BPSL1497 showing the active site disulphide bond.

Resolution (Å)	1.07
Number of reflections	49294
Protein molecules per asymmetric unit	1
Number of atoms	1024
Number of residues	105
Number of waters	179
Number of ions	4
Ramachandran favoured (%)	98.1
Ramachandran outliers (%)	0.0
RMS bond length deviation (Å)	0.008
RMS bond angle deviation (º)	1.253
Average main chain B-factors ($Å^2$)	9.0
Average side chain B-factors ($Å^2$)	13.9
Average waters B-factors ($Å^2$)	24.5
Average buffer component B-factors ($Å^2$)	11.8
R-factor (%)	0.187
R_{free} (%)	0.205
Molprobity score	1.2 (91^{st} percentile)

Table 11.5 Final refinement statistics for BPSL1497.

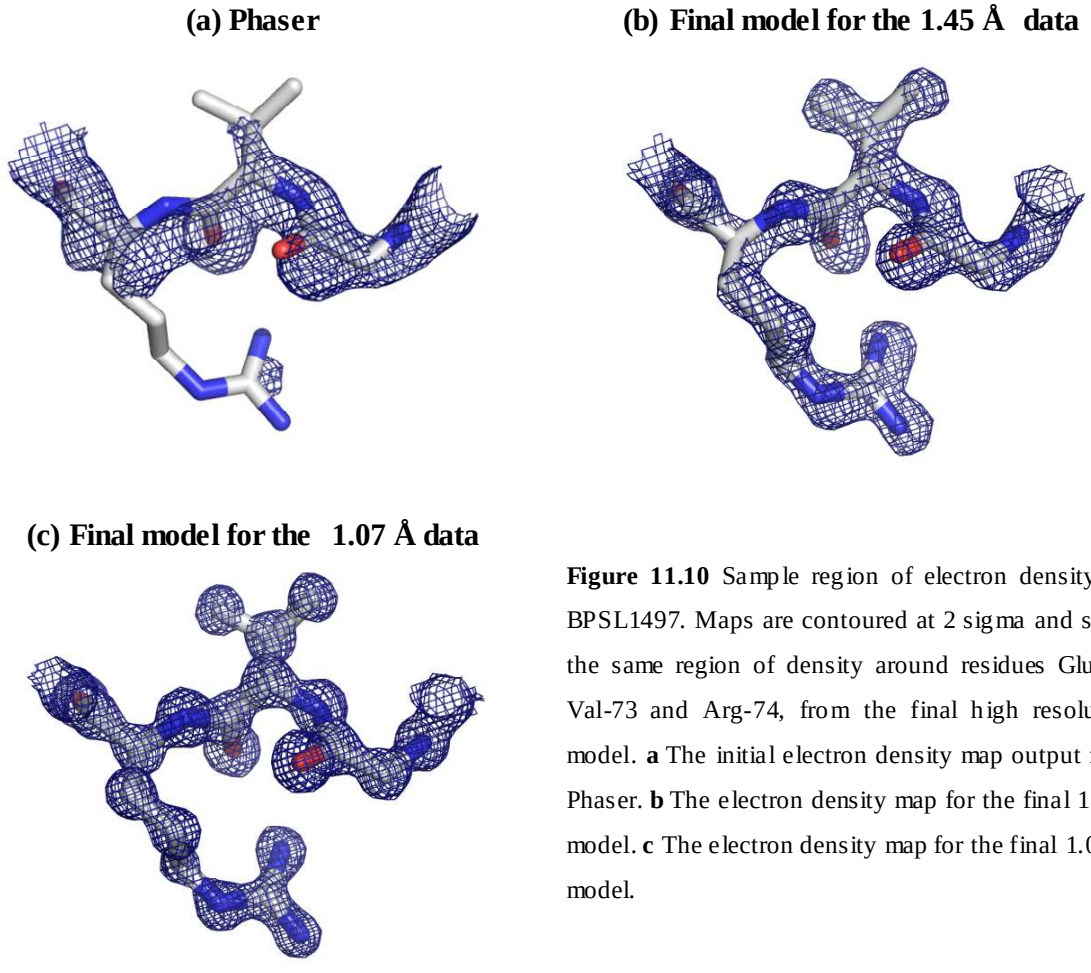
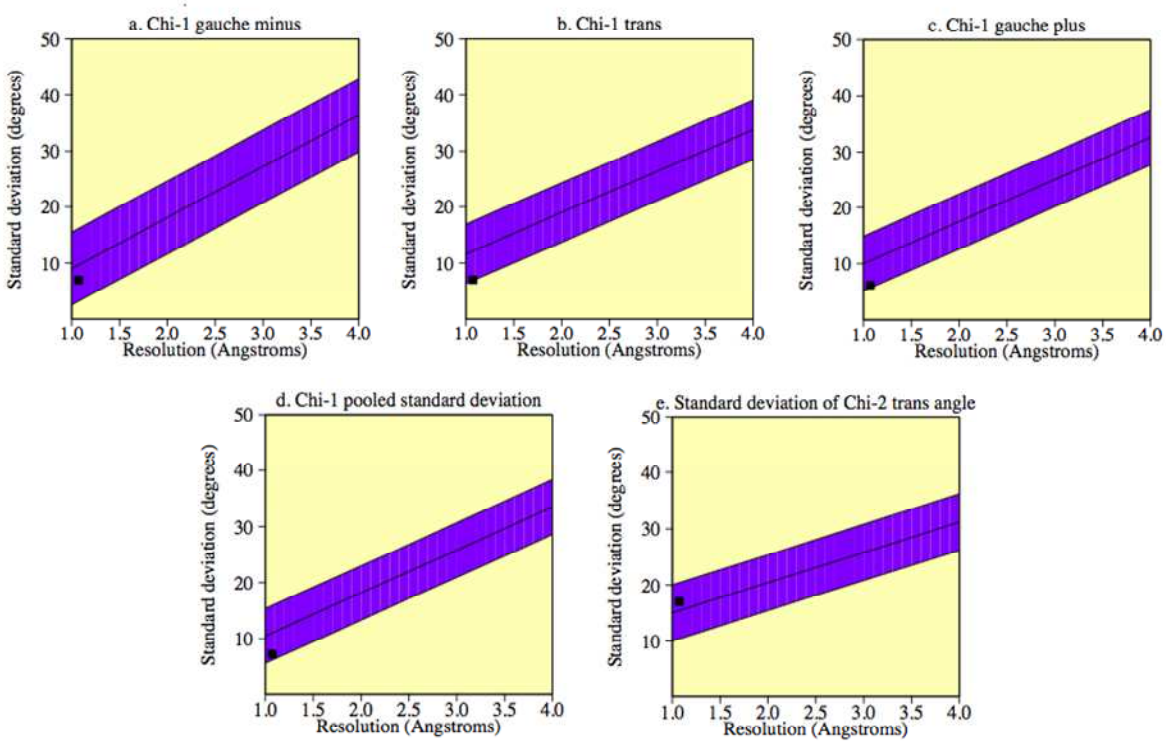


Figure 11.10 Sample region of electron density for BPSL1497. Maps are contoured at 2 sigma and show the same region of density around residues Glu-72, Val-73 and Arg-74, from the final high resolution model. **a** The initial electron density map output from Phaser. **b** The electron density map for the final 1.5 Å model. **c** The electron density map for the final 1.07 Å model.

(a) Side chain parameters



(b) Main chain parameters

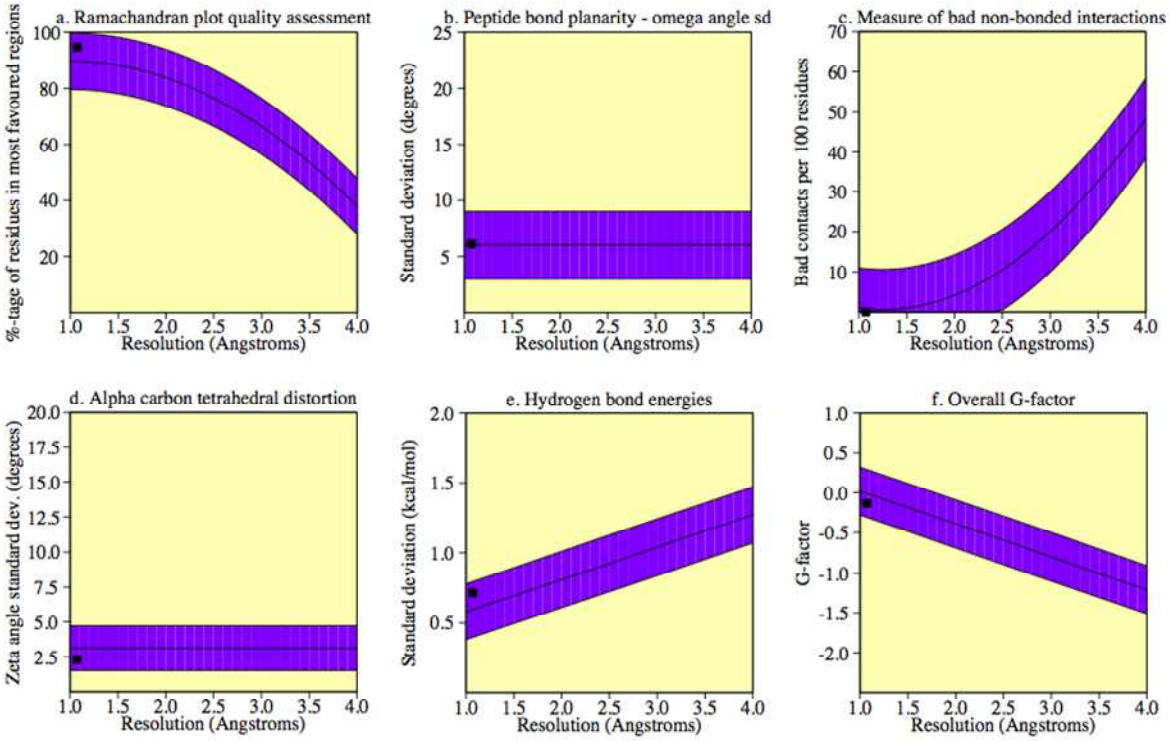


Figure 11.11 Main chain and side chain properties for the final BPSL1497 model. The figure was generated using PROCHECK [157] and shows all residues have properties within the expected range for the resolution of the data.

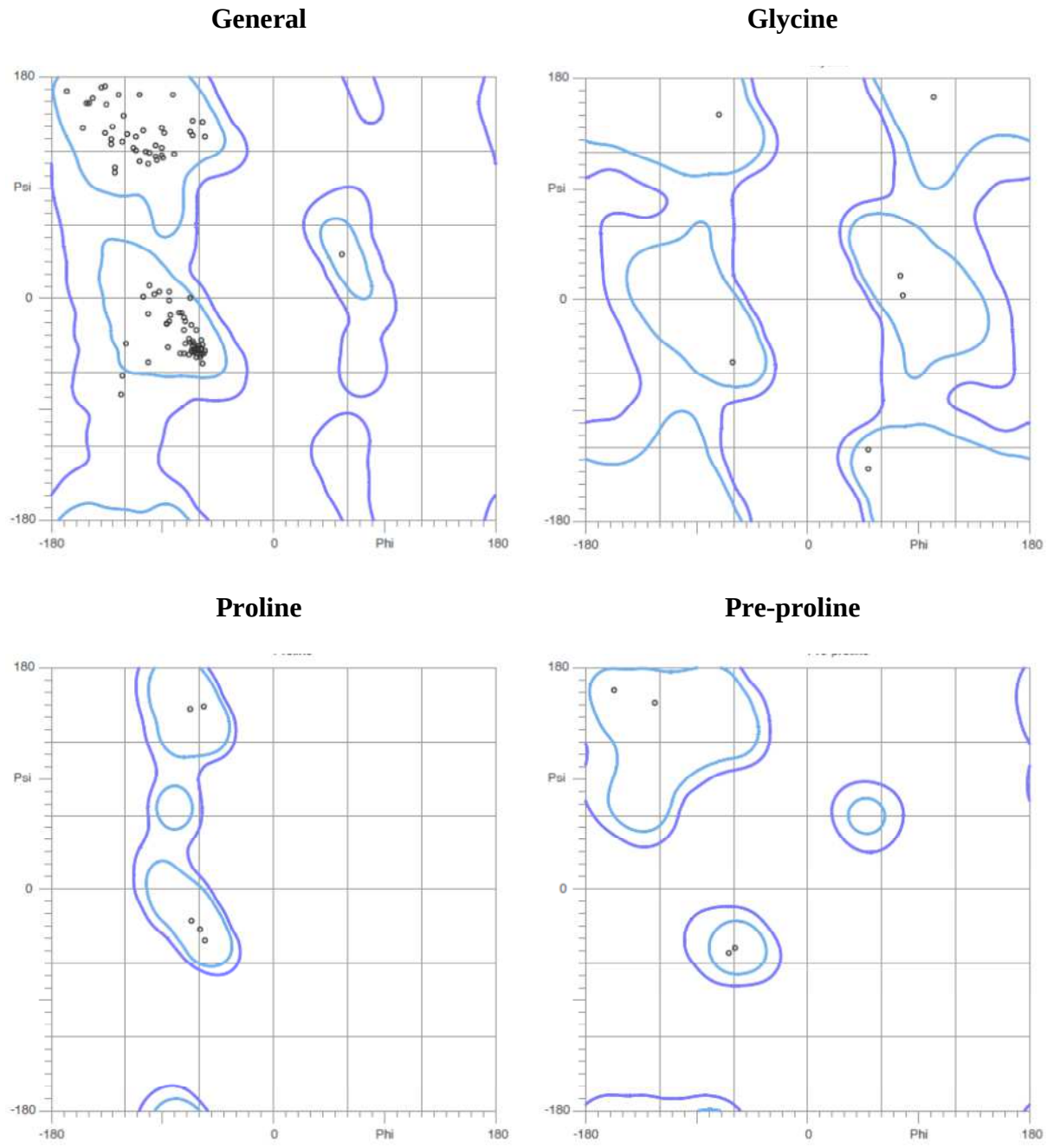


Figure 11.12 Ramachandran plot and statistics for the final BPSL1497 model. The figure was generated using Molprobity [158] and shows all residues have acceptable values for their phi and psi angles with 98.1 % falling within favoured regions.

11.9 The model of BPSL1497

The model has been restrained to standard bond lengths and angles and the root mean square deviation (RMSD) of bond lengths and bond angles in the final structure for the high resolution data is 0.008 Å and 1.253º respectively. The structure was validated using PROCHECK [160], and the Molprobity server [161] which showed the overall structure was of good quality. All main chain and side chain parameters are better than or within the expected range for the resolution of the data (figure 11.11) and all residues fell within allowed regions of the Ramachandran plot (figure 11.12).

11.9.1 Metal ions in the BPSL1497 structure

The final model contains four metal ions, which due to the presence of 200 mM Zn^{2+} in the crystallisation solution were assumed to be zinc although no analysis was done to confirm this assumption. Three of the four ions are found mediating crystal contacts between symmetry related molecules, with one of these falling on a 2-fold symmetry axis with half occupancy.

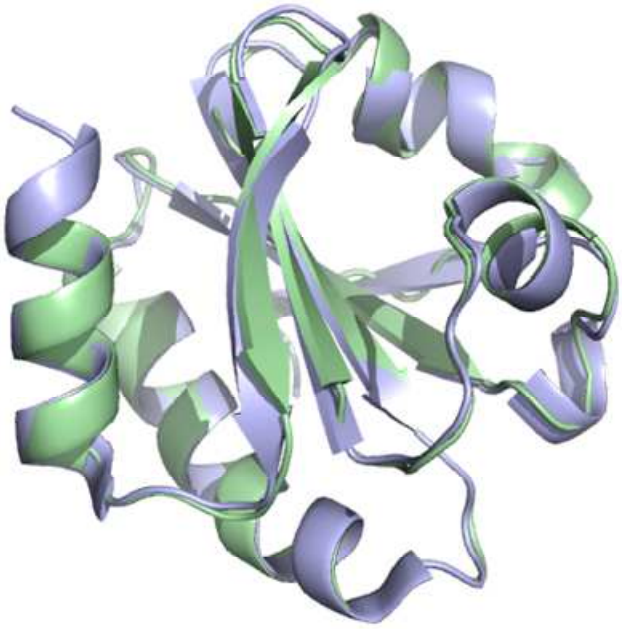
11.10 BPSL1497 is similar to thioredoxin structures from other species

As expected the overall structure of thioredoxin from *Burkholderia pseudomallei* is very similar to that from other species (figure 11.13 a). The RMSD when compared to thioredoxin from *Escherichia coli* [189] (PDB 2TRX) is 0.6 and there are only minor differences between the two structures. The active site architecture is also conserved (figure 11.3 b), as is the positioning of a number of key structural residues (data not shown) identified by sequence homology and functional studies [190].

11.10.1 Interface residues of thioredoxin and thioredoxin reductase in Burkholderia pseudomallei are conserved

In the *Escherichia coli* thioredoxin to thioredoxin reductase complex structure, the two proteins make seven hydrogen bonds with each other (table 11.6). The residues involved in these interactions are generally conserved in the *Burkholderia pseudomallei* proteins with two exceptions. In thioredoxin Tyr-70 is replaced by Phe-71 and in thioredoxin reductase Ala-237 is replaced by Gln-237 in the *Escherichia coli* and *Burkholderia pseudomallei* proteins respectively, however as these hydrogen bonds involve the carbonyl oxygens of the residues, these differences should not affect the formation of the hydrogen bonds. The positions of the residues involved in these interactions are conserved between the *Escherichia*

(a) Overall fold



(b) Active site

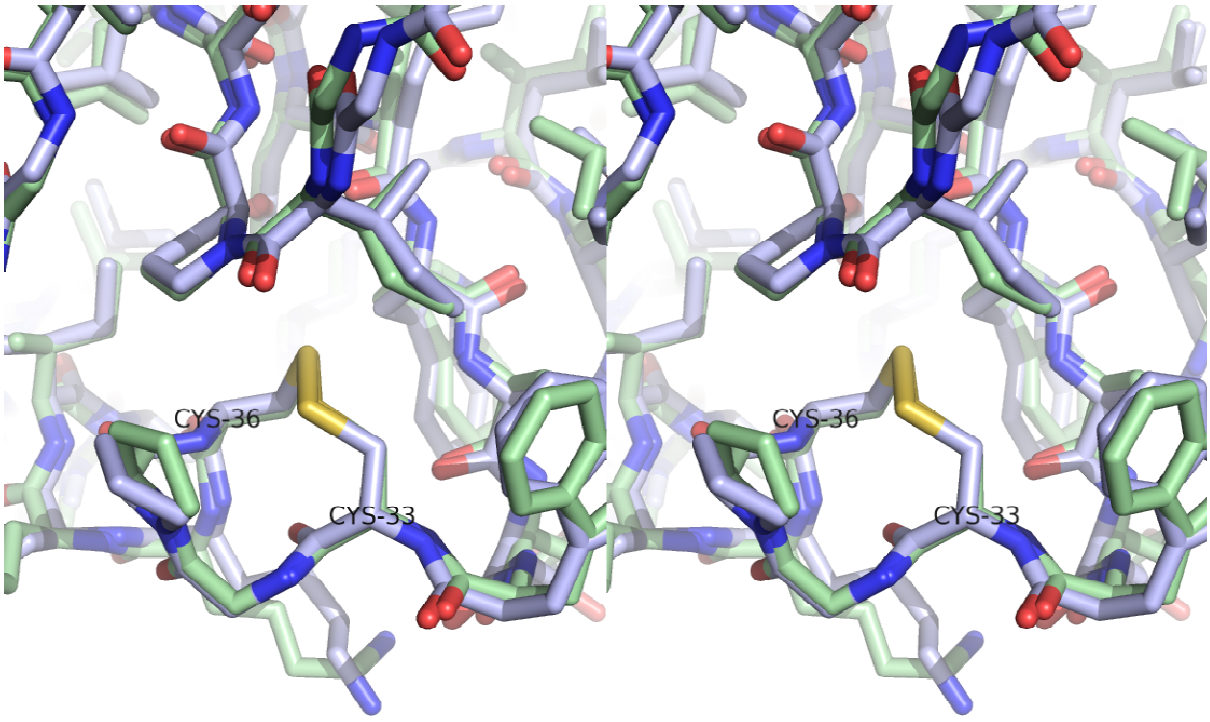


Figure 11.13 The structure of BPSL1497 is similar to other thioredoxin structures. In both images the *Burkholderia pseudomallei* structure is shown in blue and the *Escherichia coli* structure in green. **a** The overall fold of the molecules. **b** The active site architecture.

Thioredoxin residue	Thioredoxin reductase residue	Bond length (Å)
Trp-31 [NE1]	Cys-138 [CO]	3.1
Arg-73 [NH]	Gly-129 [CO]	3.2
Arg-73 [NE]	Arg-130 [CO]	3.0
Arg-73 [NH1]	Arg-130 [CO]	2.9
Arg-73 [NH1]	Ala-237 [CO]	2.7
Ile-75 [NH]	Asp-139 [OD1]	2.8
Tyr-70 [CO]	Arg-130 [NH2]	2.6

Table 11.6 Hydrogen bonds between thioredoxin and thioredoxin reductase in the *Escherichia coli* complex structure.

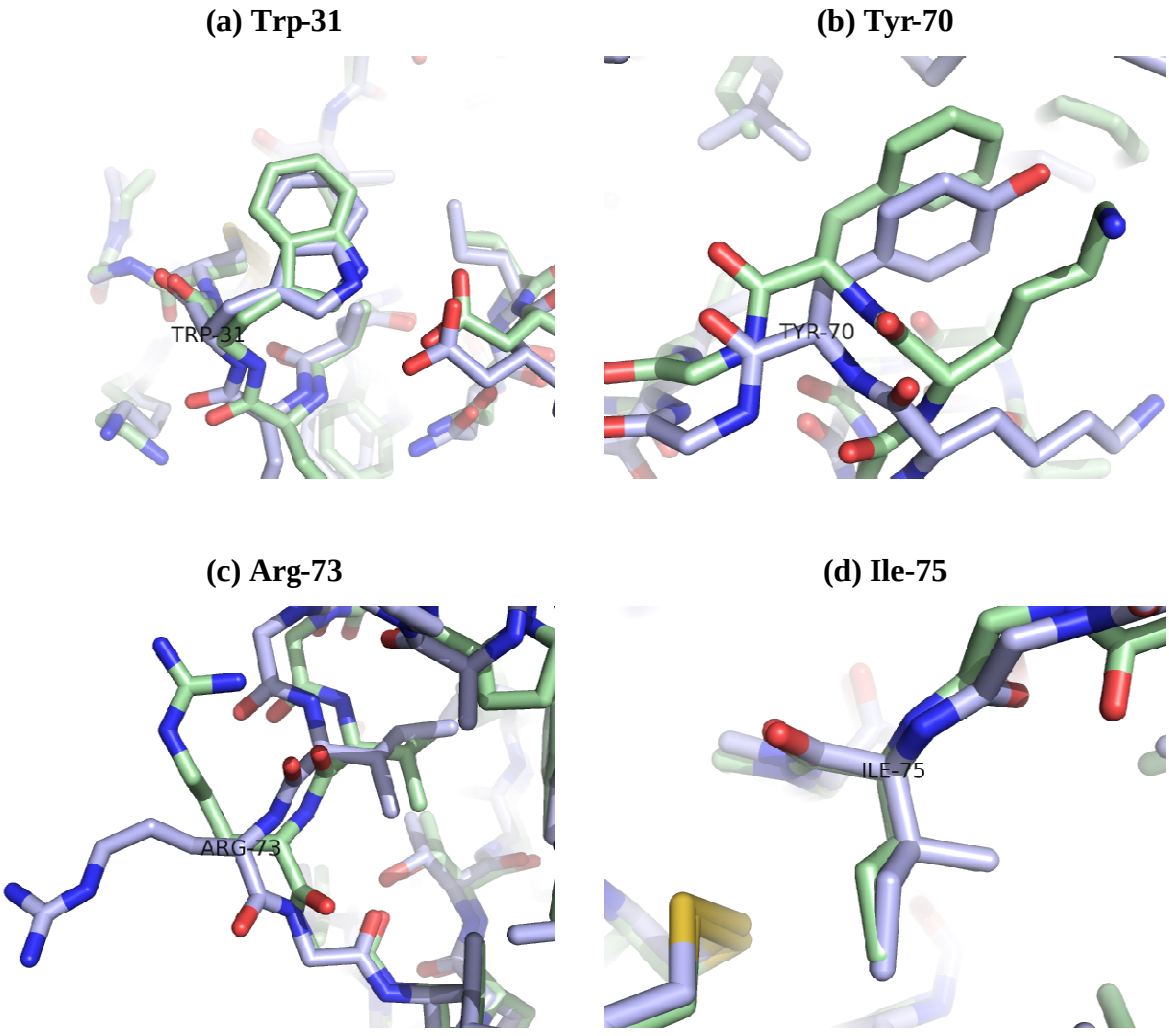


Figure 11.14 Conservation of residues forming hydrogen bonds between thioredoxin and thioredoxin reductase.

coli and *Burkholderia pseudomallei* thioredoxin structures (figure 11.13) and thioredoxin reductase structures (Martin Bush, Thesis 2009), suggesting the interactions between the two proteins in each system are very similar. This suggests the system is ideal for further studies to determine a high resolution structure of the thioredoxin reductase to thioredoxin mutant proteins covalent complex.

Chapter four

Discussion of results obtained as part of this thesis

Section 12 Summary, conclusions and future work

12.0 Summary, conclusions and future work

The work in this thesis is mainly concerned with a small-scale structural genomics project for targets from *Burkholderia pseudomallei*. An overview of the successes of the programme, together with highlights from the work on each of the selected targets, including an analysis of the structure determined for BPSS0211 are discussed in this section. A structure was also solved for thioredoxin from *Burkholderia pseudomallei* as part of an on-going project to resolve the remaining ambiguities in the thioredoxin reductase catalytic mechanism.

12.1 Overview of the structural genomics project

Nine targets were initially selected for study based on their perceived role as potential pathogenicity determinants of the human bacterial pathogen, *Burkholderia pseudomallei*. The assignments were based on a target's possession of at least one of two criteria. The first was their expression in the pathogenic *Burkholderia pseudomallei* but not the closely related, non-pathogenic *Burkholderia thailandensis*. The second required their expression be controlled as part of a stress response, which controls other known virulence factors. Once selected, the genes were cloned into expression vectors and overexpression trials were conducted. Conditions that allowed the soluble overexpression of five of the targets were found and large scale growths were carried out to obtain material from which to purify the target proteins. Each of the expressed proteins were purified and crystallised to varying degrees of success and a single structure was solved for the target BPSS0211 (table 12.1). An abundance of information has been amassed for the targets studied, however this project has been cut short by time constraints and there is a plethora of additional work to be considered, both structural and functional, that would be worthwhile conducting to extend this study.

12.2 The insoluble targets: BPSL3012, BPSS0214, BPSS1588 and BPSS2055

Sequence analysis of BPSS1588 identified an N-terminal signal sequence, directing the protein for secretion from the cell. The protein was also predicted, following secondary structure assignment and threading analysis, to possess a collagen binding domain. The four genes were cloned into expression plasmids, with the resulting sequences for BPSL3012 and BPSS1588 having no differences when compared to the K96243 genome sequence while BPSS0214 had three residue changes, L81Q, E174D and F388S, and BPSS2055 had a single change D322G. Overexpression trials yielded conditions for all four targets where the protein could be produced in the insoluble cellular fraction. However, the inability to produce soluble

Target	Cloned	Expressed	Purified	Crystallised	Structure solved
BPSL0599	Yes	Yes	Yes	Yes	No
BPSL1958	Yes	Yes	Yes	Yes	No
BPSL3012	Yes	No			
BPSS0211	Yes	Yes	Yes	Yes	Yes
BPSS0212	Yes	Yes	Yes	Yes	No
BPSS0213	Yes	Yes	Yes	Yes	No
BPSS0214	Yes	No			
BPSS1588	Yes	No			
BPSS2055	Yes	No			

Table 12.1 Summary of results for the structural genomics project.

protein for the four targets, BPSL3012, BPSS0214, BPSS1588 and BPSS2055 means regrettably further studies on these proteins would be difficult to conduct. The insolubility of the two targets that possess residue changes from the K96243 strain, BPSS0214 and BPSS2055, may be resultant from incorrect folding caused by mutations created by PCR error. The insolubility of BPSS1588 could be related to the reducing environment of the cell. As the protein contains multiple cysteine residues and usually exists outside of the cellular environment the formation of disulphide bonds might be required for correct folding. Alternative constructs and expression systems could be considered to produce soluble protein for all four targets however this would effectively require the process to be entirely restarted. Another option would be to attempt refolding experiments on overexpressed insoluble protein. This would involve a large scale growth under conditions that provide abundant amounts of insoluble protein, followed by cell lysis, protein denaturation and screening conditions to subject the proteins to for refolding. Commercial screens are available for this purpose, such as the Pierce® Protein Refolding Kit (Thermo Scientific). If solubilisation was successful, further studies could then be conducted on these targets.

12.3 BPSL0599

The gene was cloned into an expression plasmid and the resulting DNA sequence was identical to that found in the K96243 genome sequence. Overexpression of BPSL0599 resulted in two distinct molecules, one being the full length protein and the other a stable fragment consisting of the 103 N-terminal residues. Separation of these two proteins during purification was problematic and three samples were used in initial crystallisation trials. These were a mixed sample and two samples containing predominantly the full length, or the

N-terminal fragment. Conditions producing crystals from the samples consisting the two partially separated species were identified; however the crystals are currently unsuitable for X-ray analysis. Following testing, optimisation screens around conditions containing protein crystals could be conducted, or if the crystals are salt, further screening could be done. Problems in crystal optimisation might arise from the two forms of the protein that are produced during overexpression that are not completely separated during purification. As the fragments are inherently similar it is likely that the crystals might be mixtures which could be alleviated by producing a new construct corresponding to the stable fragment which could then be expressed in isolation allowing the purification of a homogeneous sample for further crystal trials.

12.4 BPSL1958

BPSL1958 contains a highly repetitive sequence of a 52 amino acids repeated six and three quarter times with the repetitive nature also apparent at the DNA level. This suggests the gene was created in a recent duplication event and there has not been enough time for the sequence of the individual repeats to drift. Analysis of the secondary structure and threading results strongly suggest this protein forms a beta propeller structure with each of the repeats forming an individual blade. The gene was cloned into an expression plasmid with a single sequence difference, S312A, compared to the K96243 genome sequence. Due to the lack of any methionine, cysteine or tyrosine residues in the primary sequence and the absence of a suitable homology model in the PDB, it was necessary to produce a number of mutants to allow phasing experiments to be conducted. Cysteine mutations (K3C, S128C, A357C, K3C-D244C, K3C-H340C) were created to allow covalent mercury binding in co-crystallisation experiments and a triple methionine mutant (K3C-I44M-I252M-I356M) was produced to allow the production and crystallisation of seleno-methionine protein. The native protein and all mutant forms of the protein were overexpressed and purified by a two-step procedure, involving anion exchange chromatography followed by gel filtration. The protein eluted from the gel filtration column anomalously with a predicted molecular weight much lower than expected, likely due to the protein interacting with the matrix of the beads. Crystallisation trials using the native protein found a single condition capable of producing crystals, following optimisation a crystal was selected and a dataset was collected. Cysteine mutant forms of the protein were all placed into co-crystallisation trials with the mercurial compound EMTS. Hits were found for all mutants and crystal optimisation was attempted. This was successful only for the K3C and S128C mutant forms with little improvement seen for the

other mutants. MAD data were collected from crystals of the K3C, S128C, A357C and K3C-H340C mutants, though the only crystals which produced any anomalous signal were obtained by a co-crystallisation experiment between the K3C mutant and EMTS. This data was used to calculate a set of initial phases for the structure although the resulting map was uninterpretable. Crystals containing the seleno-methionine K3C-I44M-I252M-I356M quadruple mutant protein were produced by seeding experiments, although the diffraction properties of these crystals and the incorporation of seleno-methionine into the protein are yet to be determined. The next step towards structure resolution for BPSL1958 would be further seeding trials for the seleno-methionine K3C-I44M-I64M-I356M mutant protein. Once useful crystals have been produced, diffraction data could be collected for this target in order to obtain a further set of initial estimates for the protein phases through Selenium MAD data collection and processing. These phases could then be combined with the phases calculated from the mercury MAD K3C mutant data. Failing optimisation of seleno-methionine mutant crystals, it would be necessary to create further mutants and subject these to purification and co-crystallisation trials. If this created an interpretable electron density map a model could be constructed through iterative rounds of model building and refinement.

12.5 BPSS0211, BPSS0212 and BPSS0213

The three genes are contained within the four gene BPSS0211-BPSS0214 operon where all four genes in the operon are annotated as having no known function. BPSS0212 and BPSS0213 are homologs of each other and are annotated as containing two domains of unknown function, the N-terminal DUF1842 and C-terminal DUF1843 which is homologous to the full length BPSS0211.

12.5.1 BPSS0211

The gene was cloned into an expression plasmid with the same sequence as found in the K96243 genome sequence. Native and seleno-methionine proteins were overexpressed and purified by a three-step procedure, involving anion exchange chromatography and an ammonium sulphate cut followed by gel filtration. The protein eluted from the gel filtration column as a shouldered peak possibly corresponding to two oligomeric states of the protein with predicted molecular weights corresponding to a dimeric and tetrameric form of the protein. Initial crystallisation screens were conducted resulting in several hits, one of which was optimised to produce the crystals used for data collection. Data was collected from both native and seleno-methionine protein crystals from the same crystallisation condition

allowing calculation of a set of initial phase estimates by MAD techniques. A 2.17 Å structure with good overall statistics was built with a single protein molecule in the asymmetric unit consisting of 51 residues arranged into two alpha helices along with some buffer components including a potentially biologically relevant Zinc ion. Analysis of the crystal packing suggested the protein arranged as a tetramer best described as a dimer of dimers. Analysis of the interfaces between subunits suggested that both dimers and tetramers were stable and the tetramer was the most probable quaternary structure. Questions remain regarding the structure of BPSS0211, ultimately higher resolution data may resolve some of the ambiguities in the model, particularly surrounding the potentially biologically relevant zinc ion. It would also be informative to determine whether residues 1-9 and 61-63 are absent or disordered in the crystal. A mass spectrometry analysis conducted on crystals of the protein would answer this question. Finally conformation of the oligomeric state of BPSS0211 in solution could be achieved by sedimentation centrifugation and native gel electrophoresis. This could be conducted in the presence and absence of Zinc and a number of other divalent metal cations to determine if their presence has an effect on the proteins quaternary structure.

12.5.2 BPSS0212

The gene was cloned into an expression plasmid with a single residue difference, Q187R, when compared to the K96243 genome sequence. Native and seleno-methionine proteins were overexpressed and purified by a three-step procedure, involving two forms of anion exchange chromatography followed by gel filtration. The protein eluted from the gel filtration column with a predicted molecular weight corresponding to a dimeric protein. Throughout the purification process, the protein was observed to be suffering from degradation which continued after purified. Crystallisation trials resulted in the production of a number of poor quality crystals that defied optimisation. Data was collected from both native and seleno-methionine protein crystals grown from the same condition allowing calculation of a set of initial phase estimates by MAD techniques. The quality of the initial phases was low and the resulting electron density map was impossible to interpret. Due to the degradation of the protein it is unclear what the contents of the crystal are, and it is likely the degradation has hampered the growth of high quality crystals. In order to produce better quality crystals needed to determine the structure it will be necessary for the crystal contents to be accurately identified by mass spectrometry or protein sequencing. The crystallised fragment could then be cloned and expressed allowing for its purification without other contaminating fragments. Alternatively it may be possible to produce a stable full length protein or to stabilise the

fragment of the protein that crystallises, long enough for optimisation trials to be conducted. The use of protease inhibitors throughout the purification process and present in the crystallisation condition might achieve this, preventing the breakdown of the protein. Once useful crystals have been produced, better quality diffraction data could be collected allowing a superior set of initial phases to be calculated. If this created an interpretable electron density map, a model could then be constructed through iterative rounds of model building and refinement.

12.5.3 BPSL0213

The gene was cloned into an expression plasmid with the same sequence as found in the K96243 genome sequence. The protein was overexpressed and purified by a two-step procedure, involving anion exchange chromatography followed by gel filtration. The protein eluted from the gel filtration column with a predicted molecular weight corresponding to a dimeric protein. Throughout the purification process, the protein was observed to be suffering from degradation which continued after purified, although this was at a much slower rate than for the related BPSS0212 protein. Crystallisation trials were conducted and a single condition has been identified that produces low quality protein crystals. The next steps towards structure resolution for BPSS0213 would be crystal optimisation trials. Once useful crystals have been produced, diffraction data could be collected for this target. The next step would be obtaining a set of initial phases, through Selenium MAD experiments. Once initial phase estimates have been produced a model could be constructed through iterative rounds of model building and refinement.

12.5.4 Understanding the role of BPSS0211 as part of the BPSS0211-BPSS0214 operon

The inferred function of BPSS0211 from the solved structure as an oligomerisation domain asks a number of questions regarding the nature of the operon it exists within. An analysis of the conservation of residues between BPSS0211 and the homologous regions of BPSS0212 and BPSS0213, and with homologs of the three proteins from other bacterial species was conducted. The conserved residues were mapped onto the BPSS0211 structure and their positions in the tertiary and quaternary structure were analysed. Conserved residues appear on one of two interfaces, either between the two helices of the monomeric structure or as part of the interface that forms the dimeric structure. This suggests the possibility of heterodimers forming between the different proteins in the operon. Determining the nature of any possible assembly of homo dimers and tetramers of the individual proteins and hetero dimers and

tetramers of the three proteins could provide an insight into the biological role of BPSS0211 within the cell. Further understanding could then be derived from ascribing a function to the other domain, DUF1842, found in BPSS0212 and BPSS0213, possibly by determining the structure of either protein.

12.6 The structural genomic approach to structure determination

The Protein Structure Initiative (PSI) funded by the American National institute of health (NIH) represents the largest structural genomics project embarked upon. The project began in 2000 with an initial pilot phase (PSI-I) focused on the demonstration of feasibility and development of methodology. This was then applied in the second phase (PSI-II) to targets on a large scale commencing in 2005. The ultimate goal of the project is to make three dimensional structural information an integral part of all biological research and to provide this information through experimentally determined structures or homology modelling for any target based on its primary sequence. To this end the programme aims to provide structural coverage for all families of proteins and therefore one selection criteria is for unique targets not represented in the PDB. For selection purposes these were defined as proteins with less than 30% identity to proteins for which structures had already been determined. Other criteria include targets of potential therapeutic interest and those required to give greater coverage of biological pathways.

One benefit of structural genomics is the generation of economies of scale in structure determination, reducing the cost of each individual structure solved. The estimated cost of solving a unique protein structure in an average structural biology group is predicted to be between $250,000 and $300,000 while the cost of solving a similar structure as part of the pilot phase of the PSI was $138,000 [191]. This figure has subsequently reduced as the project has continued and therefore in terms of structures solved structural genomics clearly represents a cost effective solution to structure determination.

The overall success rate to structure of below 3 % for individual targets selected and studied as part of phases I and II of the PSI is very low (table 12.2), although as the projects continue this is set to rise. It is unclear in large scale projects how many targets are dropped prematurely when the individual project runs into difficulties; whereas in a smaller, more focused approach, such as the one described here, more time and effort can be applied to a specific target which ultimately may result in a much higher success rate.

PSI centre	Selected	Cloned	Expressed	Purified	Crystallised	NMR	Structure in PDB
ATCG3D	7	7	7	7	6	0	6
CESG	1425	1083	932	254	71	15	40
CHTSB	129	124	29	10	0	0	5
GPCR	3	3	0	0	0	0	0
ISFI	428	330	244	210	175	0	129
JCSG	26357	25448	15667	15665	2607	0	962
MCSG	7763	6559	4759	2169	485	0	389
MPP	516	231	212	33	0	0	0
NESG	11219	4269	4064	1437	144	155	279
NYCOMPS	532	362	90	43	0	0	0
NYSGRC	9895	2391	2091	537	68	0	52
NYSGXRC	3309	2749	2373	1183	333	0	228
TMPC	97	73	53	18	0	0	0
Combined	72422	49804	34882	23205	3889	170	2090
Success from previous stage (%)		68.8	70.0	66.5	0.2		51.5
Success from target selection (%)		68.8	48.2	32.0	5.6		2.9

Table 12.2 Statistics for phase I and II of the Protein Structure Initiative.

12.6.1 Appraisal of the study described in this thesis

The approach adopted in the study can be considered a small scale, highly targeted structural genomics program. The need for new treatments for the disease melioidosis, coupled with the large proportion of genes of unknown function within the bacterium's genome, demonstrates the need for studies into the biology of *Burkholderia pseudomallei*, particularly around systems involved in the organisms pathogenicity. The ultimate aims of this project were to identify potential pathogenicity determinants, elucidate their structures with the intention of ascribing a function based on structural analysis and subsequent informed functional studies, in order to develop the understanding of the unknown aspects of the biology of *Burkholderia pseudomallei*. Improvements could be made to the target selection strategy, and the criteria for selection could be broadened. A potential expansion of the project could be to consider targets whose expression is controlled by other alternative sigma factors known to control the expression of virulence factors, such as RpoE. Selected targets could be examined in terms of the predicted success for their crystallisation, and difficult targets discarded or their sequences further analysed and modified in order to improve the likelihood of crystals being

obtained. A number of programs exist to determine these characteristics including the XtalPred server [192]. This program makes predictions based on the predicted biochemical and biophysical nature of a protein's primary sequence, such as length, pI and disordered regions [193]. The current success rate to structure of 11 % for the study described here is not particularly high, although better than for the large-scale PSI; however several of the targets are getting close to having structures determined and if given more time, the project would yield more structures and begin to shed light on the functions of the targets involved. The current lack of structural information for the remaining targets prevented the assignment of even putative functions to be tested. The only structure solved as part of this study represents a domain conserved between three out of four members of an operon. While it is predicted that the solved target, BPSS0211, plays a role in the oligomerisation of the operon components, it is unclear what role the BPSS0211-BPSS0214 plays in the organism and therefore what, if any, role it plays in pathogenicity. As such the aim of the project to further the understanding of the biology of *Burkholderia pseudomallei,* with the ultimate aim of developing new treatments, can be broadly viewed as a failure to date. Overall the project, if continued, could yield a wealth of information about the selected targets and represents a viable approach to furthering the understanding of an organism's biology that could easily be extended within *Burkholderia pseudomallei* or applied to other pathogenic bacteria.

12.7 Thioredoxin system project

The thioredoxin gene BPSL1497 was cloned into an expression plasmid which was used to overexpress the protein. This was then purified by a two-step procedure, involving anion exchange chromatography followed by gel filtration. Crystallisation screens identified a number of conditions under which the protein would crystallise in different crystal forms. One of these was optimised leading to the production of large crystals used for data collection. The crystals were of an exceptional quality diffracting to very high resolution and a number of datasets were collected. Initial protein phase estimates were calculated by molecular replacement and the structure was determined to 1.07 Å resolution. The model had reasonable statistics which could likely be improved through further refinement. The overall fold of the protein was found, as expected, to correlate well with previously solved thioredoxin structures from other species and the architecture of the active site and other important residues were also conserved.

12.7.1 Towards a structure of the FR conformation of thioredoxin reductase

The ultimate aim of this project is to determine a high resolution structure for a covalent complex between thioredoxin (BPSL1497) and the FR conformation of thioredoxin reductase (BPSL2605). Currently the only determined structure corresponding to this comes from *Escherichia coli* and the resolution only extends to 3.0 Å. At this resolution there are details that cannot be resolved, leaving unanswered questions about the mechanism of bacterial thioredoxin reductase enzymes. A similar technique to that used to produce a structure relating to the FR conformation of *Escherichia coli* thioredoxin reductase [181] could be applied to the homologous protein from *Burkholderia pseudomallei*. The *Escherichia coli* structure was solved by locking thioredoxin reductase into the FR conformation by forming a complex with thioredoxin before crystallisation. This was achieved by mutating one of the two active site cysteine residues in each protein, and forming a disulphide cross-link between the remaining cysteine residues. The thiol reagent 5,5′-dithiobis(2-nitrobenzoic acid) (DTNB) was used to "activate" the thioredoxin mutant by producing a mixed disulphide between the remaining active site cysteine and thionitrobenzoic acid. The "activated" thioredoxin mutant protein was then mixed with the thioredoxin reductase mutant protein producing the disulphide linked complex. The complex remains stable as the mutated cysteine residues are no longer able to attack the disulphide bond as would occur if the native proteins were used. The active site cysteine residues in thioredoxin and thioredoxin reductase from *Escherichia coli* are residues 32 and 35, and 135 and 138 respectively, and correspond to residues 33 and 36, and 136 and 139 of the *Burkholderia pseudomallei* proteins. For the *Escherichia coli* structure, the covalent complex was formed by linking cysteine-32 of thioredoxin to cysteine-138 of thioredoxin reductase as this is probably the physiologically relevant intermolecular disulphide [183]. Therefore the two desired mutant forms are BPSL1497 C36S and BPSL2605 C136S.

12.7.2 Obtaining ultra-high resolution data

Some of the BPSL1497 crystals tested diffracted to beyond 0.8 Å and it may therefore be possible to obtain ultra-high resolution data for this protein in the future. Ultra-high resolution data is useful in crystallography as it allows accurate experimental calculation of the geometry of proteins, which can then be used to improve all structure determination experiments. The problem with obtaining this data from the crystals of BPSL1497 is the presence of a long unit cell axis causing difficulty in separating the reflections on this axis.

One solution to collecting ultra-high resolution data from such a crystal would be to use a larger detector, allowing the collection of high angle reflections while maintaining reflection separation in the diffraction images. Higher resolution data might also be achieved by a number of data collection techniques using the available apparatus at the Diamond synchrotron light source. Testing a large number of crystals to find the least mosaic would reduce the spread of individual reflections in the images. Another technique requires the careful mounting of crystals to ensure they are rotated around the long unit cell axis during data collection. Using a smaller X-ray beam, such as the I24 microfocus beamline at Diamond which can be reduced to 5 µm by 5 µm and increasing the wavelength of the X-ray beam may also further separate the reflections.

Chapter five

Theory, materials and methods

Section 13 Cloning to purified protein

Section 14 Crystallisation to structure determination

Section 15 Abbreviations and symbols

13.0 Cloning to purified protein

The section includes a discussion of the theory behind techniques used in this thesis. It also describes in detail the laboratory techniques used to clone, overexpress and purify target proteins from *Burkholderia pseudomallei*.

13.1 Recombinant DNA technology and protein production

X-ray crystallography structural studies require large quantities of soluble, correctly folded, relatively pure protein. Proteins can be obtained from a native source although it is often unfeasible, due to low levels of expression, difficulty in obtaining enough initial material, and in this case the pathogenicity of the original organism. Recombinant DNA technologies are therefore routinely employed to produce large amounts of a target protein. The technique involves extraction of the DNA from the organism being studied followed by amplification of the gene of interest by use of PCR. The gene is then inserted into an expression plasmid and transformed into a host that can be manipulated into producing large amounts of the protein of interest.

13.1.1 Polymerase chain reaction

In a standard PCR several components are required, including a heat stable DNA polymerase enzyme, the four deoxynucleoside triphosphates (dNTPs), a DNA template containing the region that is to be amplified and DNA primers complementary to the ends of the region of template DNA that is to be amplified, all in a suitable buffer for the reaction to take place. The PCR progresses through several cycles of amplification. Each cycle consists of three phases achieved by altering the temperature of the reaction. The first phase, denaturation, consists of boiling the reaction in order to disrupt hydrogen bonds between DNA bases resulting in the production of single strands of DNA with no internal secondary structure. This is followed by the annealing phase where the temperature is dropped to below the melting point of the DNA primers allowing them to bind to their complementary sections on the template DNA. During the final elongation phase, at the polymerase's optimal temperature, the polymerase binds to the DNA and progresses along the strand producing a new copy of the desired region. The high GC content of *Burkholderia pseudomallei's* genome [49] can cause problems with strand separation and primers annealing to the genome. This is due to the increased temperatures required for strand separation and the increased frequency of primers annealing to mismatching DNA sections. The use of DMSO as a PCR additive chemically disrupts the hydrogen bonding between base pairs, facilitating DNA strand

separation and PCR efficiency, and allows the use of lower temperatures in the reaction. A lower annealing temperature can in some circumstances result in an improved amplification with either a larger yield of the desired DNA or a reduction in the number of contaminants also produced during the PCR reaction caused by primers annealing to incorrect regions of the template DNA [145].

13.1.2 pET vectors

The pET expression system (Novagen) is a range of expression vectors and hosts for the tightly controlled production of large quantities of a target protein. The plasmids used in this project, pET21a (figure 13.1) and pETBlue-1 (figure 13.2), possess a cloning region containing a T7 promoter sequence followed by a Lac operator sequence, ribosome binding site and a T7 terminator sequence. In conjunction with a suitable expression host they can be used to produce large quantities of a target protein (figure 13.3).

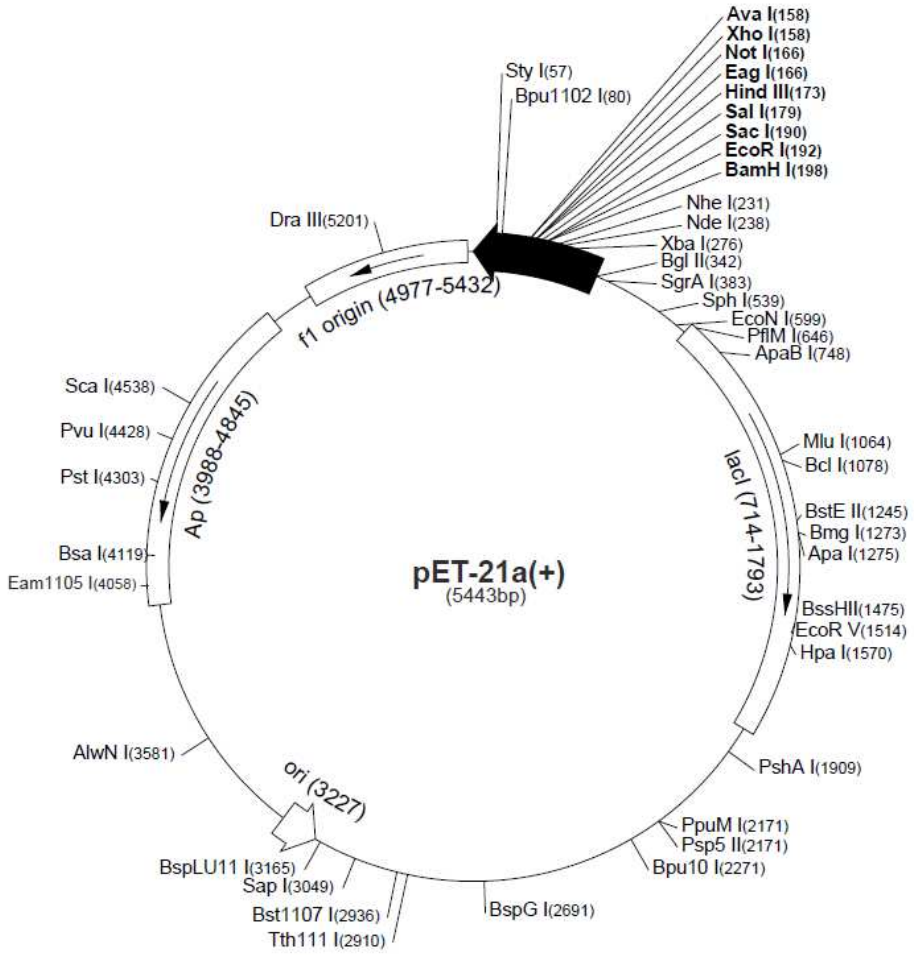
13.1.3 Cloning with pET21a

The plasmid pET21a (figure 13.1) contains multiple restriction enzyme sites in its cloning region for the insertion of the target gene by traditional restriction cloning. The restriction sites used in this project are NdeI, BamHI and EcoRI. Following restriction digestion of both the plasmid and the amplified gene fragment the insert can be ligated into the vector using T4 DNA ligase. The resulting plasmid can then be transformed into a bacterial strain lacking the T7 polymerase gene to prevent any background expression which could be problematic at this stage. The plasmid, pET21a, also contains a copy of the LacI gene to ensure sufficient repressor protein is produced for the operator sequences in the genome and the plasmid once it is transferred into an expression strain.

13.1.4 Cloning with pETBlue-1

The plasmid pETBlue-1 (figure 13.2) allows insertion of the target gene through a blunt end cloning into an EcoRV restriction site present in a copy of the α-peptide fragment of the LacZ gene. The resulting plasmid can then be transformed into a bacterial strain lacking the T7 polymerase gene to prevent any background expression and with a copy of the ω-peptide gene of LacZ in its genome. Colonies containing a plasmid with an insertion can be chosen by blue/white screening on agar containing IPTG and X-gal. The LacZ gene fragments together code for β-galactosidase which is able to break down X-gal producing a blue pigment. If the LacZ α-peptide gene is intact (no insertion) it is expressed and able to

(a) Plasmid map of pET21a

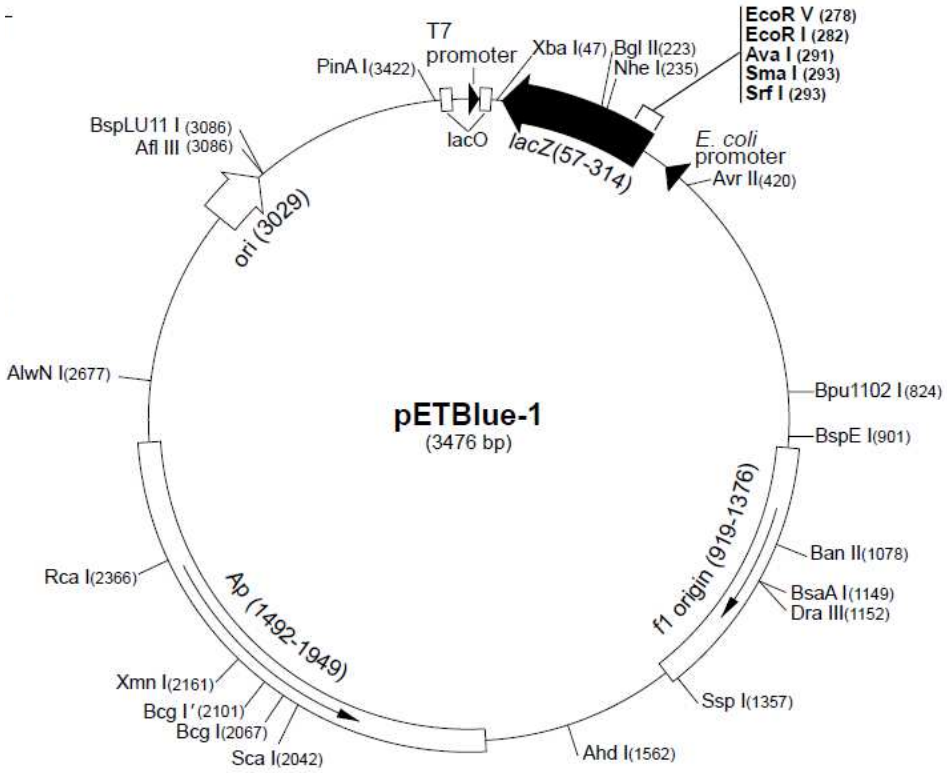


(b) Sequence of the cloning region of pET21a vector

```
BglII                T7 Promoter          lac operator          XbaI
agatctcgatcccgcgaaattaatacgactcactataggggaattgtgagcggataacaattcccctctagaaataatttgtttaactttaag

 rbs        NdeI  NheI                              BamHI       SacI       HindIII        XhoI
aaggagatatacatatggctagcatgactggtggacagcaaatgggtcgcggatccgaattcgagctccgtcgacaagcttgcggccgcactcg

                                                          EcoRI        SalI         NotI

agcaccaccaccaccaccactgagatccggctgctaacaaagcccgaaaggaagctgagttggctgctgccaccgctgagcaataactagcata
                   T7 Terminator
acccctttggggcctctaaacgggtcttgaggggttttttg
```

Figure 13.1 The pET21a plasmid. **a** The plasmid contains the cloning region (black arrow), two origins of replication from F1 phage and pBR322 and two genes; LacI, the coding sequence for the lactose operator protein and Ap, the coding sequence for β-lactamase providing ampicillin resistance. **b** Sequence of the cloning region of pET21a vector. The region contains a T7 promoter sequence followed by a lac operator sequence, ribosome binding site and multiple cloning sites.

(a) Plasmid map of pETBlue-1



(b) Sequence of the cloning region of pETBlue-1 vector

```
   T7 Promoter          lac operator          XbaI
taatacgactcactataggggaattgtgagcggataacaattcccctctagacttacaatttccattcgccattcaggctgcgcaactgttggg
                                                                    lacZ gene

aagggcgatcggtacgggcctcttcgctattacgccagcttgcgaacggtgggtgcgctgcaaggcgattaagttgggtaacgccaggattctc

                                    BglII       NheI                          rbs       EcoRV
ccagtcacgacgttgtaaaacgacggccagcgagagatcttgattggctagcagaataattttgtttaactttaagaaggagatatagatatcg

EcoRI    AvaI
aattcctgcccgggcgttgtaatcatagtcataatcaatactcctgactgcgttagcaatttaactgtgataaactaccgcattaaagctattc
                                              lacZ gene

gatgataagctgtcaaacatgataattcttgaagacgaaagggc
    lacZ gene                 rbs
```

Figure 13.2 The pETBlue-1 plasmid. **a** The plasmid contains the cloning region inside a copy of the lacZ gene (black arrow), two origins of replication from F1 phage and pBR322 and the coding sequence for β-lactamase providing ampicillin resistance. **b**. The cloning region contains a T7 promoter sequence followed by a lac operator sequence, ribosome binding site and EcoRV restriction site for blunt end cloning.

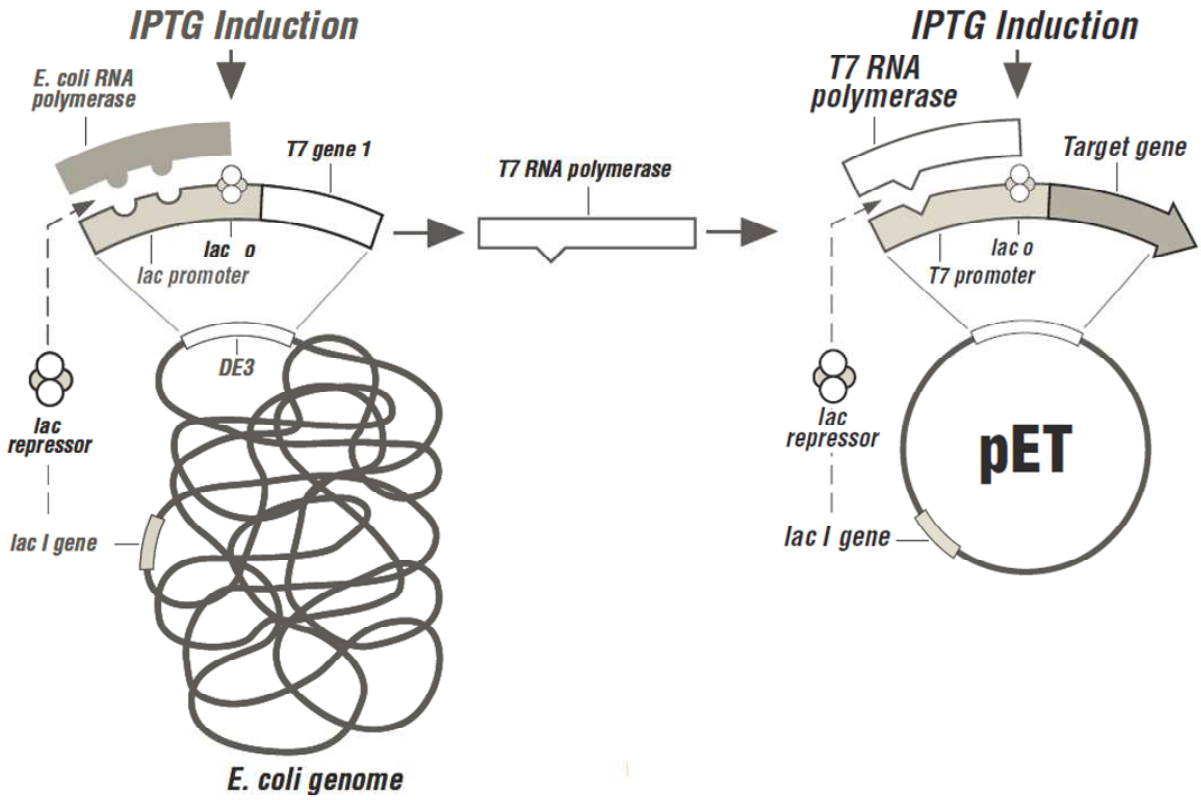


Figure 13.3 Regulating protein expression in the pET expression system. Expression hosts are all DE3 lysogen strains possessing a copy of T7 prophage DNA which contains the T7 polymerase gene downstream to a lac promoter site and lac operator binding site. They also contain a copy of the LacI gene which encodes for the lac repressor protein. Under normal conditions the absence of lactose allows the lac repressor to bind the lac operator sequence preventing expression of the T7 polymerase. When IPTG (a non-hydrolysable lactose analog) is added it binds to the lac repressor preventing it binding the operator sequence and allowing the transcription of the T7 polymerase gene. Target genes are inserted into pET vectors downstream to a T7 promoter and lac operator site. Once translated the T7 polymerase is able to transcribe a target gene inside a pET vector leading to its expression. The plasmid pET21a contains an additional copy of the lac repressor gene to ensure there is no basal expression without the addition of IPTG, pETBlue-1 does not contain a copy of the gene and must therefore be used with strains that contain an additional copy of the gene encoded on a plasmid such as pLacI strains.

combine with the ω-peptide of LacZ produced by expression from the genome resulting in a functional copy β-galactosidase and a blue colony. White colonies contain plasmids with an insertion preventing the expression of the α-peptide and therefore β-galactosidase activity. The plasmid, pETBlue-1 does not contain a copy of the LacI gene and therefore must be used in conjunction with expression strains that have a copy of this gene for protein expression such as the pLacI strains that contain a plasmid with the gene maintained by a chloramphenicol resistance marker.

13.1.5 pET expression hosts

Expression hosts in the pET expression system are all DE3 strains containing a copy of T7 RNA polymerase gene and the lacI gene within their chromosomal DNA. Expression of the polymerase is under *lacUV5* control and is induced by the addition of IPTG to a bacterial culture. Once the polymerase is expressed it can transcribe the plasmid DNA into mRNA which can then be translated, leading to production of large amounts of the target protein (figure 13.3).

13.2 Cloning into pET21a and pETBlue-1 plasmids

13.2.1 Primer design

Primers were designed for target genes using the genome sequence of the *B. pseudomallei* K96243 strain, as strain D286 which was used for PCR amplification has not been sequenced. All primers were checked to ensure both primers of each pair had a similar melting point. It was also checked that no problems would arise from self-complementarity or secondary structure formation using oligocalc [194] and *in silico* PCR was performed to ensure only one section of genomic DNA would be amplified [195]. Primers for pET21a were all between 28 and 29 nucleotides in length and contained a 6 to 8 nucleotide sequence upstream of the restriction site to allow the restriction enzymes to cleave the PCR product efficiently. Forward primers were designed to contain an NdeI (CATATG) restriction site followed by 14 to 17 nucleotides of genomic DNA starting with the second amino acid encoding codon of the gene (NdeI restriction site contains a start codon). Reverse primers contained either a BamHI (GGATCC) or EcoRI (GAATTC) restriction site, depending on the presence of either site in the gene sequence which was checked using NEBCutter [196], followed by 14 to 17 nucleotides of genomic DNA from downstream of the gene. Forward primers for pETBlue-1 cloning were designed to incorporate genomic DNA starting from the start codon of the gene while reverse primers contained genomic DNA from downstream of the gene.

13.2.2 Polymerase chain reaction amplification

Target genes were amplified from *Burkholderia pseudomallei* strain D286 genomic DNA by PCR. Reactions of a total volume of 25 µl were set up for each condition containing 1 µl genomic DNA (provided by Professor Rahmah Mohamed, University Kebangsaan, Malaysia), 10 pmol of forward and reverse primers, 12.5 µl BioMix Red (Bioline) and 0, 2.5 or 5 % (v/v) DMSO. The reaction mixes were subjected to an initial denaturation at 95 ºC for

5 minutes before thirty cycles of denaturation at 95 ºC for 30 seconds, annealing at 58 ºC for 30 seconds and elongation at 72 ºC for 60 seconds before a final elongation at 72 ºC for 5 minutes.

13.2.3 Polymerase chain reaction product purification

PCR products were run on a 1 % (w/v) TAE agarose gel containing GelRed at 100 V for 45 minutes to allow separation of DNA. Bands were visualised on a UV light box and those of the correct length were removed from the gel using a scalpel. DNA was extracted from the gel using a QiaQuick® Gel Extraction Kit (Qiagen) and the standard protocol. All centrifugation was done at 17,000 g in a benchtop centrifuge for 60 seconds. Gel slices were weighed in mg and the weight was taken to be equivalent to 1 gel volume in µl. The pieces were first suspended in 3 volumes of buffer QG and incubated at 50 ºC for 10 minutes. Once the gel pieces had dissolved 1 volume of isopropanol was added to the mixture before it was applied to a QIAquick® column. The column was centrifuged and the flow-through was discarded before adding 500 µl buffer QG, centrifuging and again discarding the flow-through. A further wash was carried out using 750 µl buffer PE, centrifuging, discarding the flow-through and again centrifuging to ensure all ethanol had been removed from the column. The column was then transferred to a clean microcentrifuge tube and the DNA was eluted by adding 65 µl water, incubating at room temperature for 2 minutes, and centrifuging. Yields varied between 20 and 50 ng μl^{-1}.

13.2.4 Vector production for pET21a cloning

A colony of DH5α *Escherichia coli* cells containing plasmid vector pET21a growing on LB agar containing 100 µg ml^{-1} carbenicillin was picked into 3 ml LB medium containing 100 µg ml^{-1} ampicillin and grown overnight at 37 ºC at 200 rpm. Cells were then harvested by centrifugation at 5,000 g for 20 minutes before the plasmids were extracted using a QIAprep Spin MiniPrep Kit (Qiagen) and the standard protocol. Cell pellets were resuspended in 250 µl buffer P1; this buffer contains RNase to degrade any RNA preventing it contaminating the purified plasmid. The cells were then subjected to alkaline lysis by the addition of 250 µl buffer P2 and mixing by gentle inversion. The solution was neutralised and the salt concentration adjusted to allow binding to a QIAprep spin column by the addition of 350 µl buffer N3. DNA adsorbs to the silica membrane of the spin column in the presence of high concentrations of chaotropic salts at pH below 7.5. The lysate is then cleared of cell debris by centrifugation at 17,000 g for 10 minutes in a benchtop centrifuge. All subsequent

centrifugation was done at 17,000 g for 60 seconds. The supernatant was applied to a QIAprep spin column, centrifuged and the flow-through discarded. A wash was carried out using 750 µl buffer PE, which contains ethanol to precipitate the DNA on the column while removing the salts. The column was centrifuged, the flow-through discarded and the column was again centrifuged to ensure all ethanol had been removed from the column. The column was then transferred to a clean microcentrifuge tube and the DNA was eluted by adding 65 µl of water, incubating at room temperature for 2 minutes and finally centrifuging to collect the pure plasmid DNA. Yields from minipreps typically vary between 10 and 100 ng μl^{-1}.

13.2.5 Restriction digestion, ligation and transformation for pET21a cloning

Reactions were set up containing 65 µl purified vector or insert DNA, 40 Units NdeI (NEB) and 40 Units BamHI (NEB) or EcoRI (NEB) in 1 x NEB buffer 3 (NEB). Reactions were incubated at 37 ºC for 6 hours before increasing the temperature to 65 ºC for 30 minutes to denature the restriction enzymes. Digested vector DNA was purified by gel extraction using a QiaQuick® Gel Extraction Kit (Qiagen) and the protocol as described previously. Digested PCR products were recovered using a QiaQuick® PCR clean up Kit (Qiagen). All centrifugation was done at 17,000 g for 60 seconds. Reaction mixtures were mixed with 5 volumes buffer PB and applied to a QIAprep spin column, centrifuged and the flow-through discarded. 750 µl buffer PE were added to the column which was centrifuged, the flow-through discarded and the column was again centrifuged to ensure all ethanol had been removed from the column. The column was then transferred to a clean microcentrifuge tube and the DNA was eluted by adding 65 µl of water, incubating at room temperature for 2 minutes and finally centrifuging to collect the pure digested plasmid DNA. Ligation reactions were set up containing 20 units T4 DNA ligae (NEB), 50 ng digested vector and a 3 x molar excess of digested insert in 1 x NEB ligation buffer (NEB) and were incubated at 16 ºC overnight. Ligation reactions were then used to transform DH5α *E. coli* cells. Eppendorf tubes containing 50 µl aliquots of competent cells were removed from the -80 ºC freezer and incubated on ice for 5 minutes. Once defrosted 5 µl of the ligation reaction was added before being left to incubate on ice for 30 minutes. The cells were subjected to a heat shock at 42 ºC for 30 seconds before being returned to ice for 2 minutes. 500 µl of SOC media was then added and the cells were then incubated at 37 ºC at 200 rpm for 60 minutes before plating on LB agar containing 100 $\mu g\ ml^{-1}$ carbenicillin to select for transformants.

13.2.6 Ligation and transformation for pETBlue-1 cloning

Ligation reactions were set up containing 2 µl purified PCR product (approximately 50 ng) and 50 ng AccepTor vector in Clonables^{TM} ligation premix buffer (Novagen) and were incubated at 16 ºC for 30 minutes. Ligation reactions were then used to transform Novablue *E. coli* cells. Eppendorf tubes containing 50 µl aliquots of cells were removed from the -80 ºC freezer and incubated on ice for 5 minutes. Once defrosted 1 µl of the ligation reaction was added before being left to incubate on ice for 5 minutes. The cells were subjected to a heat shock at 42 ºC for 30 seconds before being returned to ice for 2 minutes. 250 µl of SOC media was then added and the cells were then incubated at 37 ºC at 200 rpm for 60 minutes before plating on LB agar containing 100 µg ml^{-1} carbenicillin, 15 µg ml^{-1} tetracycline, 70 µg ml^{-1} X-gal and 80 µM IPTG for selection and blue-white screening of colonies to select for transformants producing white colonies containing an insert in the vector.

13.2.7 Confirmation of cloning results

For both pET21a and pETBlue-1 cloning, colonies were picked into 100 µl water and boiled at 100 ºC for 10 minutes. Cell debris was removed by centrifugation at 17,000 g for 5 minutes. Reactions of a total volume of 10 µl were set up for each condition containing 1 µl of cell extract, 5 µl BioMix Red (Bioline) and 5 pmol of each of a pair of relevant primers. To confirm results from pET21a cloning a single reaction using T7F (TAATACGACTCACTATAGGG) and T7R (GCTAGTTATTGCTCAGCGG) primers was conducted for each colony. Results from pETBlue-1 cloning were confirmed using a pair of reactions for each colony using either bpsl1497F and pETBlueDOWN or pETBlueUP and bpsl1497R primers. The reaction mixes were subjected to an initial denaturation at 95 ºC for 1 minute before thirty cycles of denaturation at 95 ºC for 30 seconds, annealing at 58 ºC for 30 seconds and elongation at 72 ºC for 60 seconds before a final elongation at 72 ºC for 5 minutes. PCR products were analysed by electrophoresis on a 1 % (w/v) TAE agarose. colonies producing the correct sized band(s) were picked into 3 ml LB containing 100 µg ml^{-1} ampicillin and grown overnight at 37 ºC at 200 rpm. The cells were harvested and plasmids were purified using a QIAprep Spin MiniPrep Kit (Qiagen) and the same protocol as before. Purified plasmids were sent for sequencing using T7F and T7R primers (SourceBioscience or Geneservice).

13.2.8 Site directed mutagenesis

Mutations were created in genes using a QuikChange site-directed mutagenesis kit (Stratagene). The strategy involves a PCR amplification of a template plasmid containing the gene of interest using mutagenic primers before a DpnI restriction digest and transformation into competent cells. Template plasmid DNA must come from a dam^+ strain of bacteria to ensure it is methylated as DpnI is a restriction enzyme that is specific for methylated DNA. Therefore the template DNA will be digested leaving behind the newly synthesised mutated plasmids. Primers were designed to be between 24 and 45 base pairs in length with the desired mutation roughly in the centre and a melting temperature above 78 ºC with a GC content above 40 % and terminating in a C or G base. PCR reactions of a total volume of 25 µl were set up for each mutation containing 5 ng template DNA, 75 ng of forward and reverse primers, 0.5 µl dNTP mix and 0.5 µl (1.25 U) PfuTurbo DNA polymerase in reaction buffer. The reaction mixes were subjected to an initial denaturation at 95 ºC for 5 minutes before eighteen cycles of denaturation at 95 ºC for 30 seconds, annealing at 55 ºC for 30 seconds and elongation at 68 ºC for 10 minutes before a final elongation at 68 ºC for 10 minutes. 5 µl of the PCR was taken for analysis by agarose gel electrophoresis, 0.5 µl (5 U) DpnI was added to the remaining 20 µl and it was incubated at 37 ºC for 2 hours. 5 µl of the DpnI digest was taken for analysis by agarose gel electrophoresis and 5 µl was taken for transformation into XL1-Blue competent cells using the same protocol as for DH5α cells. Transformants were selected by plating on LB agar containing 100 µg ml^{-1} ampicillin and grown overnight at 37 ºC before being picked into LB containing 100 µg ml^{-1} ampicillin and grown again overnight at 37 ºC at 200 rpm. The plasmids were recovered using a QIAprep Spin MiniPrep Kit (QIAGEN) and the same protocol as before. Purified plasmids were sent for sequencing using T7F and T7R primers (SourceBioscience or Geneservice) to confirm the mutations had been achieved.

13.3 Protein overexpression

13.3.1 Transformation

Plasmids, containing the correct inserts, were transformed into BL21 (DE3), Tuner (DE3) or Tuner (DE3) pLacI competent *Escherichia coli* cells (Novagen) for overexpression. Eppendorf tubes containing 20 µl aliquots of cells were removed from the -80 ºC freezer and incubated on ice for 5 minutes. Once defrosted 1 µl of plasmid DNA (approximately 10 ng $µl^{-1}$) was added and the reaction was left to incubate on ice for 15 minutes. The cells were

subjected to a heat shock at 42 ºC for 30 seconds before being returned to ice for 2 minutes. 500 µl of SOC media was then added and the cells were incubated at 37 ºC at 200 rpm for 30 minutes before plating on LB agar containing 100 µg ml^{-1} carbenicillin for BL21 (DE3) and Tuner (DE3) cells or 100 µg ml^{-1} carbenicillin and 35 µg ml^{-1} chloramphenicol for Tuner (DE3) pLacI cells to select for successful transformants.

13.3.2 Small-scale overexpression trials

An initial overnight culture grown in LB containing 50 µg ml^{-1} ampicillin for BL21 (DE3) and Tuner (DE3) cells or 50 µg ml^{-1} ampicillin and 35 µg ml^{-1} chloramphenicol for Tuner (DE3) pLacI cells was used to provide a 2 % (v/v) inoculation of 500 ml LB containing 50 µg ml^{-1} ampicillin for BL21 (DE3) and Tuner (DE3) cells or 50 µg ml^{-1} ampicillin and 35 µg ml^{-1} chloramphenicol for Tuner (DE3) pLacI cells. The cultures were grown at 37 ºC and 200 rpm until the optical density at 600 nm reached 0.8 at which point the culture was split into 50 ml aliquots in 250 ml conical flasks. Each individual culture was induced by the addition of IPTG and subjected to a different post-induction environment. Conditions tested to provide a soluble overexpressed protein were temperature (4 – 37 ºC), time (1 – 30 hours) and IPTG concentration (0.1 – 1 mM). Prior to induction and throughout the trials 1.5 ml samples were taken and the cells were pelleted by centrifugation at 17,000 g for 5 minutes in a bench top centrifuge. Samples were stored at -20 ºC before cell lysis using BugBuster® (Novagen) and PAGE analysis. Cell pellets were resuspended in 100 µl BugBuster® A with the addition of 1 µl BugBuster® B and incubated at room temperature for 15 minutes. Cell debris and insoluble protein were removed by centrifugation at 17,000 g for 10 minutes and the supernatant was taken as the soluble fraction. The pellet was resuspended in 100 µl 4 % (w/v) SDS and incubated at room temperature for 15 minutes. The sample was centrifuged at 17,000 g for 10 minutes and the supernatant was taken as the insoluble fraction. The overexpression was then analysed by running fractions from the various conditions on a suitable polyacrylamide gel to determine the optimum post-induction conditions.

13.3.3 Large scale overexpression

Once successful conditions had been found for each target the growth was scaled up to between 2 and 4 l of media in 500 ml aliquots in 2 l conical flasks. Following overexpression cells were harvested by centrifugation at 5,000 g for 45 minutes before resuspension in approximately 100 ml LB. The cells were split between 50 ml Falcon tubes and repelleted by

centrifugation at 5,000 g for 30 minutes. The supernatant was discarded and the cell pellets were stored at -20 ºC until required.

13.3.4 Production of seleno-L-methionine incorporated proteins

An initial overnight culture grown in LB containing 50 µg ml^{-1} ampicillin was used to provide a 2 % (v/v) inoculation of 2.5 l LB containing 50 µg ml^{-1} ampicillin in 500 ml aliquots within 2 l conical flasks. Cultures were grown at 37 ºC and 200 rpm until the optical density at 600 nm reached 0.6 at which point the cells were harvested by centrifugation at 5,000 g for 45 minutes before resuspension in approximately 100 ml minimal media. Minimum medium contains K_2HPO_4 10.5 g l^{-1}, $(NH_4)_2SO_4$ 1 g l^{-1}, KH_2PO_4 4.5 g l^{-1}, $Na_3C_6H_5O_7 \cdot 2H_2O$ 500 mg l^{-1}, Glycerol 5.0 g l^{-1}, Adenine 500 mg l^{-1}, Guanosine 500 mg l^{-1}, Thymine 500 mg l^{-1}, Uracil 500 mg l^{-1}, $MgSO_4 \cdot 7H_2O$ 1.0 g l^{-1}, Thiamine 4 g l^{-1}, L-lysine 100 mg l^{-1}, L-phenylalanine 100 mg l^{-1}, L-threonine 100 mg l^{-1}, L-isoleucine 50 mg l^{-1}, L-leucine 50 mg l^{-1} and L-valine 50 mg l^{-1}. The cells were repelleted by centrifugation at 5,000 g for 30 minutes before being resuspended in 2.5 l minimal medium containing 40 mg l^{-1} seleno-L-methionine and split between five 2 l conical flasks. The cells were incubated at 37 ºC and 200 rpm until the optical density at 600 nm reached 0.8 at which point the cells were induced with IPTG under the same conditions as for native protein production but with the time of overexpression doubled to allow for the slower production of protein. Following overexpression the cells were harvested and stored as for native proteins.

13.4 Protein purification techniques

X-ray crystallography requires large amounts of pure (preferably greater than 95%), homogeneous (not suffering from degradation), concentrated (typically between 5 and 30 mg ml^{-1}) protein. In order to obtain a suitable purity a number of techniques can be used to separate the protein of interest from contaminating proteins and molecules. The theory behind the purification techniques used in this thesis is described here with the details for each individual target found in the relevant results section.

13.4.1 Cell disruption

For all targets the first step in the purification was to lyse the cells containing the desired protein. This was achieved by the process of sonication which employs the use of high frequency sonic pulses to break down the bacterial cell walls and membranes. Cell pellets were removed from the freezer, defrosted and resuspended in an appropriate buffer before

homogenization by sonication using three 16 micron pulses of 20 seconds on, 20 seconds off. The crude cell extract was cleared of cell debris and insoluble protein by centrifugation at 70,000 g for 15 minutes before purification.

13.4.2 Ion exchange chromatography

The net surface charge of a protein can be used to separate it from a mixture by using its relative affinity for charged groups displayed on the surface of beads. The surface charge of a protein is dependent upon the amino acids present on its surface and the pH of its environment. If the pH of a solution is at the isoelectric point (pI) of a protein there is no overall charge on its surface. At a pH below the pI of a protein the net surface charge will be positive and at a pH above the pI of a protein the net surface charge will be negative. Ion exchange chromatography can be split into two main types; cation exchange chromatography uses negatively charged functional groups to bind positively charged cations while anion exchange chromatography uses positively charged functional groups to bind negatively charged anions. Protein solutions are loaded onto a column in a low molarity buffer allowing oppositely charged proteins to adsorb to the beads. The ionic strength of the buffer is then increased by the addition of sodium chloride which competes for the charged groups causing proteins to desorb and elute from the column with the more highly charged proteins eluting later than those with a lesser charge (figure 14.4). DEAE sepharose is an example of anion exchange chromatography and consists of a matrix of cross-linked sepharose forming beads with DEAE exposed on its surface. This technique is used as an initial purification step for all the target proteins in this thesis.

13.4.3 Ammonium sulphate cut

The solubility of a protein is a function of its concentration, and the pH and ionic strength of the solution of its environment. The surface of a protein contains hydrophobic patches caused by the presence of exposed hydrophobic residues. The addition of anti-chaotropic salts, such as ammonium sulphate, to a solution will cause a protein to precipitate once a threshold concentration is reached as water is removed from the cages around these patches which can then associate with each other. Using this technique a protein can be purified by first precipitating contaminants that drop out of solution at a lower salt concentration threshold which can be removed by centrifugation. The supernatant can then be taken and the salt concentration further increased to above the threshold for the target protein. This can then be

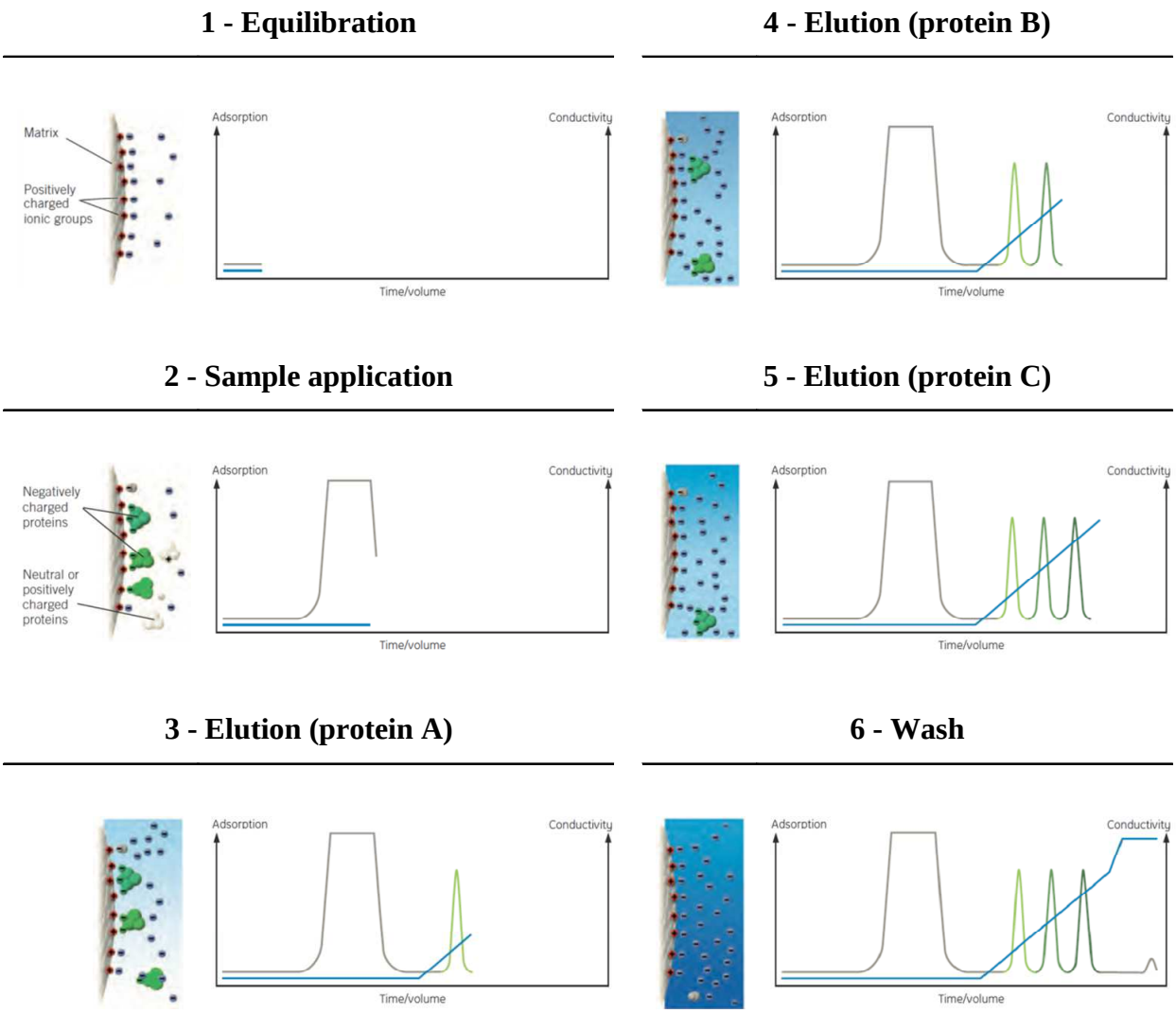


Figure 13.4 Anion exchange chromatography. Schematic representation of the separation of three proteins from a mixture displaying different charge properties. Protein A (light green) is the least negatively charged, protein B (intermediate green) is more negatively charged and protein C (dark green) is the most negatively charged that interacts with the beads. In the sample there are also other contaminants that are even less negatively charged and do not bind the column (grey). Traces show conductivity to follow salt concentration (blue) and UV absorption to follow protein elution (grey (for equilibration and sample application), light green (for protein A), intermediate green (for protein B), and dark green (for protein C)). First the column is equilibrated in a low salt buffer (1) before the sample is applied (2). Once this has run through the column a gradient of increasing salt concentration is applied (3-5) causing proteins to elute once a threshold level is reached. Finally the column is washed with a high salt buffer to remove any strongly bound contaminants (6). Figure adapted from GE Healthcare manual: Ion exchange chromatography and chromatofocusing, principles and methods.

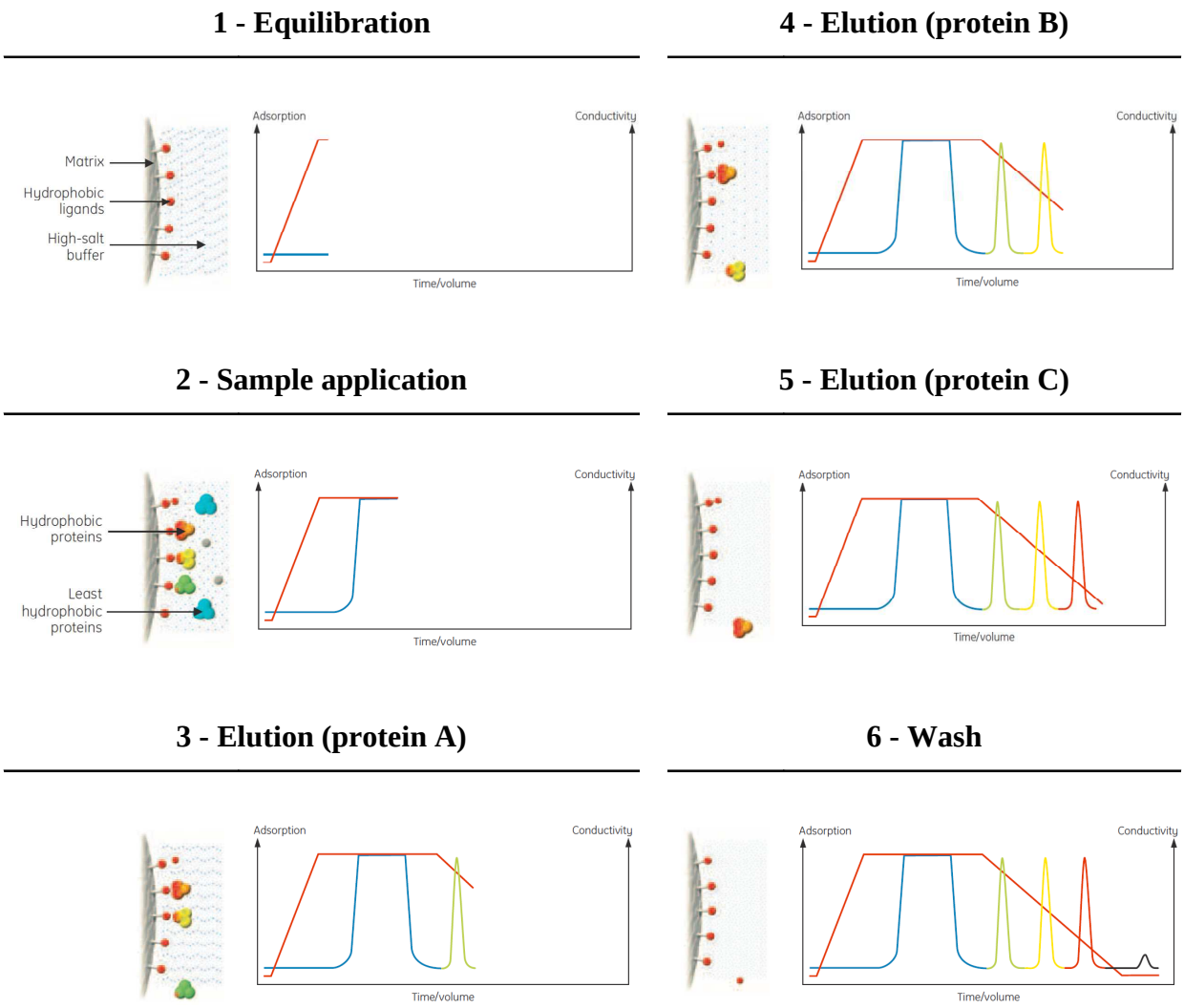


Figure 13.5 Hydrophobic chromatography. Schematic representation of the separation of three proteins from a mixture displaying different hydrophobicity properties. Protein A (green) is the least hydrophobic, Protein B (yellow) is of intermediate hydrophobicity and protein C (orange) is the most hydrophobic. In the sample there are also other contaminants that are even less hydrophobic and do not bind the column (blue). Traces show conductivity to follow salt concentration (red) and UV absorption to follow protein elution (blue (for equilibration and sample application), green (for protein A), yellow (for protein B), and orange (for protein C)). First the column is equilibrated in a high salt buffer (1) before the sample is applied (2). Once this has run through the column a gradient of decreasing salt concentration is applied (3-5) causing proteins to elute once a threshold level is reached. Finally the column is washed with water to remove any strongly bound contaminants (6). Figure adapted from GE Healthcare manual: Hydrophobic interaction and reversed phase chromatography, principles and methods.

pelleted by centrifugation and the pellet resuspended in a low molarity buffer leaving further contaminants in the supernatant.

13.4.4 Hydrophobic chromatography

Differences is surface hydrophobicity can be exploited to separate a mixture of proteins by using their relative affinity for hydrophobic groups. The presence of anti-chaotropic salts allow hydrophobic patches on the surface of a protein to interact with hydrophobic groups attached to a matrix by competing for solvation. Protein samples are loaded onto a column in a high molarity buffer allowing hydrophobic patches on proteins surfaces to interact with hydrophobic groups on the beads resulting in adsorption. The molarity of the buffer is then decreased causing proteins to desorb and elute from the column with the more hydrophobic proteins eluting last (figure 14.5). Phenyl toyopearl is an example of hydrophobic chromatography and consists of a matrix of cross-linked toyopearl forming beads with phenyl groups exposed on its surface.

13.4.5 Size exclusion chromatography

Gel filtration is a technique that separates a mixture of proteins based on their size and shape by passing them through a column filled with porous beads. The volume inside the column is in two parts, the excluded volume outside the beads, and included volume inside the beads. Molecules larger than the pores are unable to enter the beads and can therefore only occupy the excluded volume whereas molecules that are small enough to enter the beads are able to occupy both the included and excluded volumes. Therefore larger molecules will elute first from the column, and as the size of a molecule decreases it can enter a larger proportion of the beads, retarding its progress down the column and causing it to elute later, until finally proteins that can enter all beads elute last (figure 14.6). Assuming a protein is globular there is a linear relationship between the elution volume of a particular molecule from a given column and the logarithmic value of its molecular weight based on its partition coefficent (equation 13.1).

$$K_{av} = \frac{Eluted\ volume - Void\ voume}{Total\ volume - Void\ volume} \qquad \text{(equation 13.1)}$$

Therefore gel filtration can also be used to estimate the molecular weight of a protein by comparing its partition coefficient, K_{av}, to a calibration curve of partition coefficient plotted against the log of molecular weight for a particular gel filtration column determining its oligomeric state. In this thesis two separate gel filtration columns were used, a 120 ml Superose 6 column (figure 13.7a) and a 120 ml Superdex 200 column (figure 13.7b)

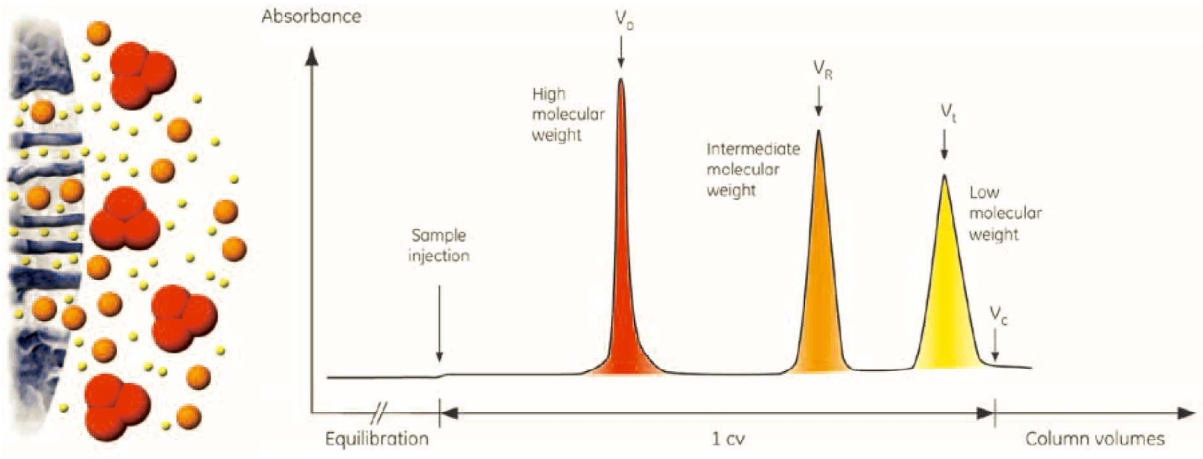


Figure 13.6 Size exclusion chromatography. The diagram shows the separation of three proteins of different sizes, including the ability of them to enter beads due to size and an elution profile. The large protein (red) is unable to enter any beads and elutes first in the excluded volume. The small protein (yellow) is able to enter all the beads and elutes at the end of the total volume. The intermediate protein (orange) is able to enter some but not all beads and elutes between the large and small proteins. Figure adapted from GE Healthcare manual: Gel filtration, principles and methods.

13.4.6 Protein concentration

Concentrated protein solutions are required for use in crystallographic studies and throughout the purification process it is often necessary to reduce the volume of a protein sample before application to a column. Concentration was performed using Vivaspin sample concentrators (GE Healthcare) which consist of two chambers separated by a polyethersulfone membrane containing pores of a controlled size. Concentration is achieved by forcing a solution through the membrane by centrifugation, allowing solvent and small molecules to pass through, leaving larger molecules trapped above the membrane. Concentrators were also used for exchanging buffers prior to crystallisation trials by a series of concentrations and dilutions

with the desired buffer. Vivaspin concentrators are available in a range of sizes to deal with volumes between 20 ml and 50 µl and molecular weight cut offs from 100,000 to 3,000 Da.

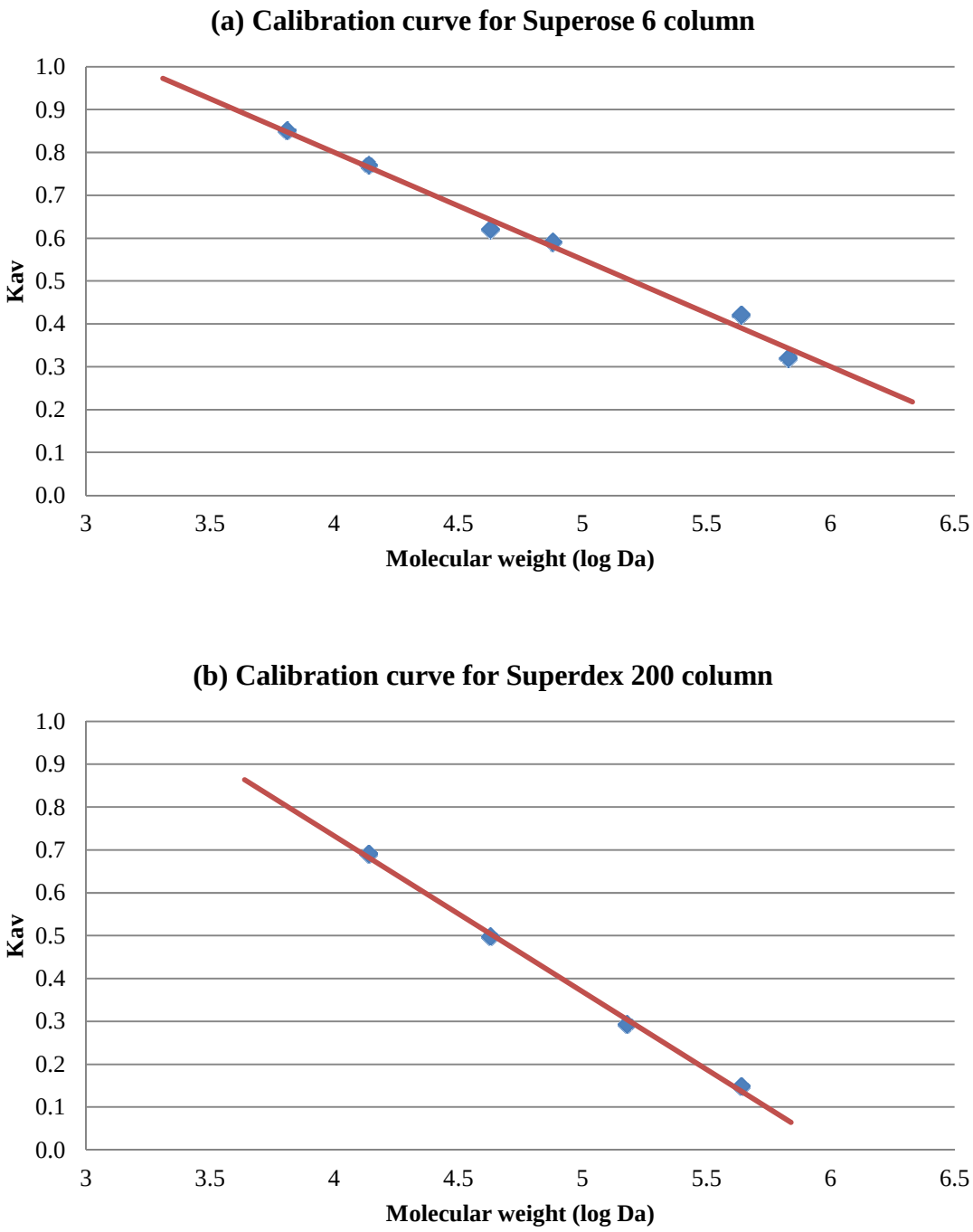


Figure 13.7 Calibration curves for the gel filtration columns used in this study.

14.0 Crystallisation to structure determination

The section includes a discussion of the theory behind techniques used in this thesis. It also describes in detail the laboratory techniques used to crystallise and solve the structures for target proteins from *Burkholderia pseudomallei.*

14.1 Crystals, space-groups and symmetry

Crystals are orderly arrays of individual molecules packed together to form a repeating lattice. In protein crystals, the molecules are held together by non-covalent interactions. These typically consist mainly of hydrogen bonding networks, involving the protein residues and ordered water molecules inside the crystal, and some ionic interfaces between protein residues and charged buffer components. Each identical repeating unit can contain multiple copies of the protein and is termed a unit cell. They consist of a volume defined by three lengths, a, b and c, and three angles, α, β and γ. The arrangement of unit cells within the crystal and any internal symmetry within the unit cell can be described in terms of the crystals space-group. There are a total of 230 possible space-groups for simple non-chiral molecules created using combinations of rotational, translational, inversion and reflectional symmetry operators. The chiral nature of protein molecules limits the functions that can be used to create symmetry elements to rotations and translations leaving 65 possible space-groups found in protein crystals. The space-group is defined by its lattice type, providing the arrangement of lattice points in three dimensions, and crystal system which determines symmetry operators inside the unit cell. There are 14 Bravais lattice types made up of the 7 crystal systems (triclinic, monoclinic, orthorhombic, tetragonal, trignal, hexagonal and cubic) determined by the unit cell lengths and angles, coupled with the five lattice centering operations (primitive, centred, face centred, internal and rhomohedral). There are only four possible rotational symmetries allowed in three dimensions, producing 2-fold (180º), 3-fold (120º), 4-fold (90º) and 6-fold (60º) related molecules, as three dimensional space is limited to these packing constraints. There are a further eleven possible screw axes, where rotational symmetry is coupled with a translation along the rotation axis. These symmetry elements are represented by the numbers following the letter denoting lattice type, with each axis with a symmetry element greater than 1 listed in order. The four rotational symmetries are designated 2, 3, 4 and 6 with the possible screw axes of 2_1, 3_1, 3_2, 4_1, 4_2, 4_3, 6_1, 6_2, 6_3, 6_4, and 6_5. The asymmetric unit (ASU) is the smallest unit within the crystal that contains all the information, that when coupled with the symmetry operations defined as part of the space group, is able to reproduce the entire crystal lattice structure. A unit cell may also contain

additional symmetry elements that are not defined by the space group. Non-crystallographic symmetry arises when there are additional symmetry operators present between molecules in the ASU that are not aligned with the other operators of the unit cell.

14.2 Producing protein crystals

Small molecules are relatively easy to crystallise, however, obtaining protein crystals is a complex and time consuming process. The difficulty in forming protein crystals is due to their size, complexity, surface mobility, complex electrostatic nature and the chemical and physical stability of protein molecules. The process requires the use of chemical precipitants, these are molecules that are good at driving the precipitation of proteins without denaturing them. Common precipitants include salts, high molecular weight straight chain polymers (such as PEGs) and organic solvents. The wide range of protein characteristics make it impossible to predict suitable conditions for any given protein therefore requiring screening of a large array of possible conditions. There are many variables for crystallisation trials including some commonly changed parameters, such as the components, their concentration and the overall pH of the crystallisation solution, alongside the protein concentration and temperature of crystallisation experiment. There are also other factors that are harder to control, such as vibration and sound, convection rates, protein sample age and consistency with previously used samples, and general cleanliness of equipment.

The process of protein crystallisation is a phase transition phenomenon which can be represented by a phase diagram (figure 14.1). The experiment is initially set up with a protein in an aqueous solution alongside other precipitant molecules at concentrations unable to drive protein precipitation. As the experiment progresses, protein and precipitant concentrations increase as water is lost, typically through evaporation or dialysis, ultimately pushing the protein out of solution as either ordered crystals or amorphous precipitant. The first stage in crystal growth is achieved once concentrations reach the nucleation phase during which small clusters of precipitated protein form reducing the overall concentrations. Crystal growth then occurs spontaneously once the concentration of protein and precipitants are at the reduced supersaturated level that allows the protein to precipitate at a rate which allows individual protein molecules to aggregate together in an orderly manner. In the ideal situation the protein solution will reach the nucleation stage and then quickly enter the growth stage preventing further nucleation leading to a smaller number of larger crystals. Crystal growth

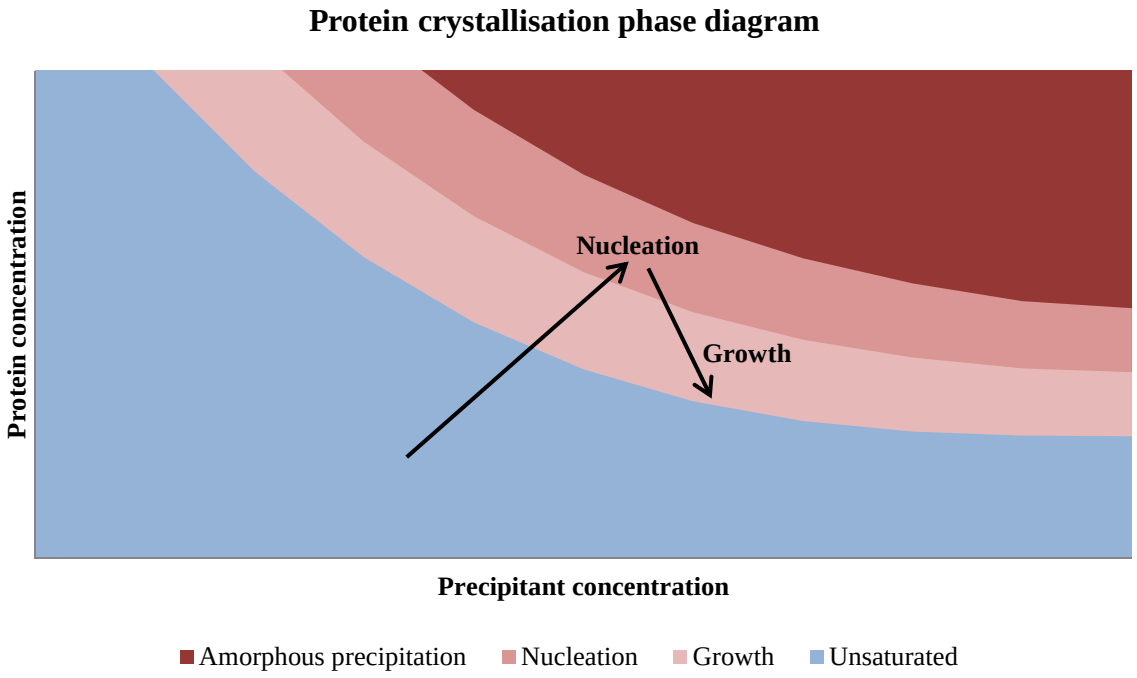


Figure 14.1 Protein crystallisation phase diagram. The blue zone represents an unsaturated solution where crystallisation will never occur. The three red zones represent increasingly unstable supersaturated solutions with deepening shade, where protein molecules can be spontaneously driven out of solution.

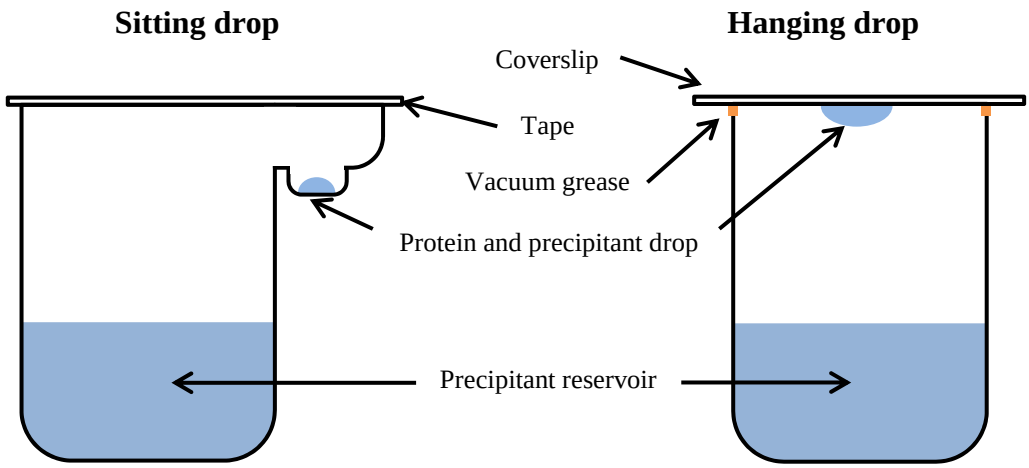


Figure 14.2 Vapour diffusion crystallisation techniques. In the sitting drop technique, as used for initial robot screening, the drop containing a mixture of protein and precipitant solutions is placed in a small well on a shelf above a well containing a large volume of precipitant solution before the system is sealed by the addition of clear tape. In the hanging drop technique, as used for optimisation screening, the drop containing a mixture of protein and precipitant solutions is suspended on the underside of a siliconised coverslip which is placed on top of a well containing a large volume of precipitant solution with vacuum grease providing an airtight seal to the system.

ceases once either the concentration of protein is no longer supersaturated or the growing crystal faces are poisoned effectively ending the ordered array. Poisoning can occur by a protein molecule being added in the incorrect orientation, the addition of a damaged protein molecule or the incorporation of a different molecule.

14.2.1 Vapour diffusion

The most commonly used method and the one used in this thesis for the production of protein crystals is vapour diffusion. The method involves creating a drop of protein solution with a precipitant concentration lower than that of a reservoir in a sealed system. Evaporation can then occur allowing diffusion of water between the drop containing the protein and the large reservoir of precipitant solution increasing the concentration of protein and precipitant in the drop until equilibrium is reached. There are two main methods of setting up vapour diffusion experiments, using either a hanging drop or a sitting drop technique (figure 14.2).

14.2.2 Robot screens

Initial screens to find conditions that yield crystals were conducted using a Matrix Hydra II PlusOne robot (Thermo Scientific). This allowed a large number of conditions to be tested quickly using relatively small volumes of purified protein solution. Screens containing 96 different conditions (Qiagen) were put down in sitting drop plates. Details of the screens conducted for each target can be found in the relevant results section.

14.2.3 Optimisation

The parameters of the initial crystallisation hit can be manipulated in an attempt to produce larger better quality crystals. Parameters of the experiment that can be changed include protein concentration, precipitant concentration, size of drops, ratio of precipitant to protein solution in drops, pH, temperature and addition of additives etc. Optimisation experiments were conducted by hand using the hanging drop technique. Again details for individual targets can be found in the relevant sections.

14.3 Principles of diffraction

Diffraction of electromagnetic radiation is dependent upon both the size of the object and the wavelength of the radiation. In order for diffraction to occur the wavelength of the light must be smaller than the distance between the objects one wishes to resolve. As the aim of X-ray crystallography is to produce an atomic model of a proteins structure, electromagnetic

radiation corresponding to inter-atomic distances, typically between 1.5 and 3.5 Å, must be used. Electromagnetic radiation of this type falls within the X-ray range, 0.1 – 100 Å, of wavelengths.

14.3.1 Diffraction from crystals

The internal symmetry and repeating nature of crystals leads to the production of distinct patterns of spots in a diffraction image. The unit cell of a crystal can be divided up by planes that cut across the unit cell axes defined by Miller indices. These have three components, h, k, l and determine the number of times the planes intersect the unit cell axes a, b and c respectively. Bragg's law (equation 14.1) describes the conditions under which a set of these planes inside a crystal produce a reflection in the diffraction image (figure 14.3).

$$2d_{hkl}sin\theta = n\lambda \qquad \text{(equation 14.1)}$$

The law states that when the angle of incidence, θ, has the properties that the difference in distance travelled by X-rays diffracting from successive planes of given spacing, d_{hkl}, is equal to an integral number of the X-rays wavelength, λ, the diffracted X-rays will be in phase and interfere constructively producing a spot, or reflection, in the diffraction image. Under conditions that do not satisfy Bragg's law, diffracted X-rays are not in phase with each other and therefore cancel each other out by destructive interference.

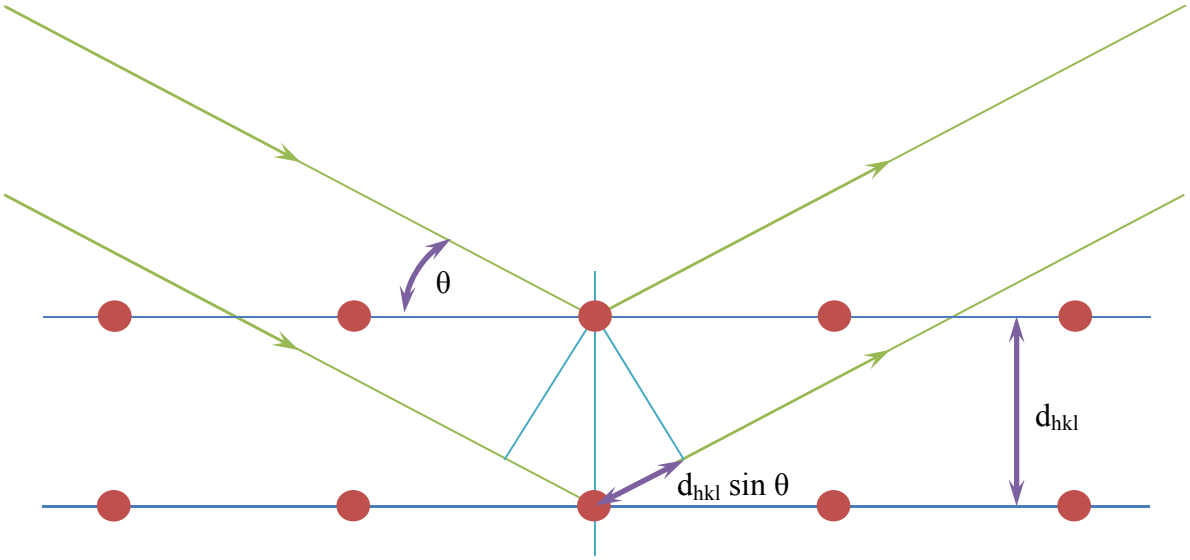


Figure 14.3 Schematic representation of Bragg's law.

14.3.2 Mosaicity

The mosaicity of a crystal is a measure of the misalignment between domains containing several unit cells in the crystal. The degree of mosaicity found in diffraction experiments is also dependent upon the X-ray beam parameters. A perfect crystal in which all the unit cells are perfectly aligned, positioned in a point source X-ray beam, would have a mosaicity value of zero and the spots in a diffraction image would be infinitesimally narrow. Crystals with a high degree of mosaicity, typically greater than 1.5º, are useless for data collection. The resultant diffraction images from a highly mosaic crystal contain smeary reflections that bleed into each other preventing the accurate assignment of an intensity to the reflection.

14.3.3 Analysis of diffraction data

The diffraction pattern can be used to determine the unit cell lengths and angles, and the space group, including the presence of rotational symmetry elements, and screw axes. The unit cell lengths are inversely proportional to the spacing between reflections in the diffraction pattern and can therefore be directly measured. There are in total 11 Laue groups that describe the symmetry found in crystal diffraction patterns. Friedel's law states that because of the centrosymmetry of the reciprocal lattice, centrosymmetrically related pairs of structure factors, h, k, l and -h, -k, -l, possess identical amplitudes but opposite phases as they originate from opposite sides of the same crystal plane. Therefore all diffraction data possess an intrinsic symmetry (in the absence of anomalous diffraction) with further symmetry between reflections indicating the presence of rotational symmetry within the unit cell. Screw axes can be determined by the presence of systematic absences, repeating patterns along an edge where one term has an intensity of zero. As the hand of the screw axis is not resolved it is not possible to distinguish between the four enantiomorphic pairs (3_1 and 3_2, 4_1 and 4_3, 6_1 and 6_5, 6_2 and 6_4) of screw axis as they produce an identical pattern of systematic absences.

14.4 Data collection apparatus

Diffraction images are collected by placing crystals in an X-ray beam and rotating them so that all the diffraction data can be collected, using a detector, from the different planes present within the crystal.

14.4.1 X-ray sources

In X-ray crystallography two distinct techniques are commonly used to produce X-ray radiation. The first method involves bombarding a rotating metal anode with electrons.

Electrons are released from a heated filament and accelerated towards a metal anode inside an electric field. These high energy electrons are capable of colliding with and displacing electrons in the target atoms. If an electron is displaced from a low lying orbital, an electron from a higher energy orbital will drop down, emitting a photon as it does. The wavelength of the resultant electromagnetic radiation is dependent on the metal it is emitted from, as a result of unique energy spacing between electron shells for different elements. For protein crystallography the commonest metal target is made of copper and emits two distinct energies of radiation. CuKα radaition is produced when electrons drop from the L to the K shell and has a wavelength of 1.54 Å, while CuKβ radaition is produced when electrons drop from the M to the K shell and has a wavelength of 1.39 Å. For crystallography a single wavelength of radiation is required so the CuKβ wavelength is removed by absorption by nickel or through the use of a monochromater. Initial crystal testing and some data collection for this thesis, was conducted at Sheffield, using a Rigaku MM007 copper rotating anode generator.

Synchotron light sources represent the second common source of X-ray radiation used in protein crystallography. A synchotron consists of a particle accelerator that injects electrons into circular particle storage ring. Inside the ring electrons travel at close to the speed of light driven by radio transmitters and maintained in a circular motion by powerful magnets. Electrons forced to follow a curved path are constantly changing velocity and therefore emit radiation at a tangent to the curve. The intensity of the resultant radiation is increased by the presence of two forms of accessory device. Wigglers and undulators increase the intensity of the emitted radiation by changing the direction of the electrons multiple times over a short distance. Beamlines are placed tangentially around the storage ring at intervals providing X-ray beams in which crystals can be mounted in for data collection. The principal advantages of synchrotron lightsources are the intensity of the radiation allowing data to be collected over much shorter exposure times, and the ability to select the wavelength of the X-ray beam therefore allowing anomalous dispersion phasing experiments to be conducted. Data collection for this thesis was primarily conducted at the Diamond synchrotron lightsource in Oxford on the I02, I03, I04 and I24 beamlines.

14.4.2 Detectors

Once the diffracted pattern of X-rays has been produced it must be recorded to be of any use; this is achieved through the use of detectors. The intensity and direction of the diffracted X-rays can be recorded, but unfortunately the phase information is lost and must be

subsequently determined using other methods (section 14.10). Image plate detectors consist of a sheet of plastic coated in thin layer of a phosphor, typically crystals of BaFEu. X-rays are able to interact with the Eu^{2+} causing it to become excited and lose an electron producing Eu^{3+}. Following exposure the image plate is read using a fine red laser (λ = 633 nm), which causes the transition of Eu^{3+} back to Eu^{2+} and the resultant emission of light (λ = 390 nm) which can be detected by a photomultiplier. Initial crystal testing and some data collection for this thesis, was conducted at Sheffield, using a MarResearch MAR345 image plate detector system. The detector systems on the beamlines at the Diamond synchrotron lightsource are not image plates but charged-coupled device (CCD) detectors or Pilatus detectors. CCD detectors consist of a phosphor screen linked to a CCD by fibre optic cables. Diffracted X-rays strike the phosphor and are converted into photons of visible light that are directed towards a pixel on CCD sensor chip through the fibre optic cable. Following exposure the pixels are read and the data is recorded. The Pilatus system is a single photon counting pixel array where the incoming photons are directly measured allowing data to be collected continuously. The principle advantage of the Pilatus system over CCD detectors, and CCD detectors over image plate detectors, is the reduced read-out time, reducing the data collection experiment time and therefore the amount of radiation damage sustained by the sample. The Pilatus detector can also produce a greater signal to noise ratio by applying an energy threshold to the photons that are counted.

14.5 Cryoprotection and radiation damage

Protein crystals typically contain only 25 to 70 % protein by volume, with the remaining space made up from solvent filled channels running throughout the crystal. Coupled with the weak nature of the interactions holding the lattice together, often including water molecules and other buffer components, this means crystals are fragile and must be kept hydrated at all times. During data collection protein crystals are susceptible to damage from the X-ray radiation used in the experiment. This can either be direct damage from the interaction of X-ray photons with atoms in the protein structure, or indirect damage from reactive radical species created following interaction of an X-ray photon with a buffer component. In order to reduce the effect of indirect damage, data is collected from samples kept at 100º K, effectively preventing the diffusion of reactive radicals through the crystal. The temperature is achieved by placing the crystal in a stream of gaseous nitrogen on the collection apparatus, flash freezing the crystal and surrounding buffer. Prior to freezing the crystal must be placed in a cryoprotectant solution, containing anti-freeze agents, in order to reduce the formation of

ice crystals. If ice crystals are present they produce a distinctive pattern of rings in the diffraction images, potentially obscuring reflections and can also potentially disrupts the crystal lattice. In order to collect the required diffraction data from a crystal it must be supported and rotated about an axis perpendicular to the incident X-ray beam.

14.6 Crystal mounting

Crystal manipulation is typically carried out using thin nylon loops, with diameters appropriate to the crystal size, attached to metal pins on magnetic bases that can be easily mounted onto the data collection apparatus. Crystals are removed from a drop by sliding the loop into the solution beneath the crystal and lifting with an upward movement resulting in the crystal being held by surface tension in a small volume of the drop solution inside the loop. Crystals are typically removed from the crystallisation solution to another drop, for soaking experiments or cryoprotection, before looping for attachment to the data collection apparatus and freezing in the cryostream. The magnetic base of the pin is attached to the goniometer, which allows the crystal to be rotated during data collection and for the precise alignment of the crystal in the X-ray beam.

14.7 Data collection

A preliminary analysis of the protein crystal properties is conducted by collecting a small number of diffraction images, with a small rotation, from different angles, typically two or three images with ϕ start angles of 0º, 45º and 90º. These images can be indexed to allow the space group, unit cell dimensions and mosaicity to be determined (section 14.3) and the overall quality of diffraction to be judged. A data collection strategy is then developed to allow all the required data for a complete dataset to be collected taking these factors into account.

14.7.1 Data collection strategy variables

The three main variables to decide are the degrees of data collected, the increment angle and the X-ray dose. The required wedge of diffraction data is determined by Friedel's law and the overall symmetry of the crystal. As a result of Friedel's law the maximum overall rotation of the crystal in the beam (for a non-anomalous dispersion experiment) is 180º in order to detect all the unique reflections. The space group of the unit cell further reduces the required angle based on which of the 11 Laue groups to which they belong. However it is usually advantageous to collect a larger dataset, to increase redundancy in the data, unless the sample

begins to suffer from radiation damage. The increment in the rotation angle over which images are recorded is decided upon based on the level of mosaicity in the crystal and the lengths of the unit cell axes. A crystal with a high degree of mosaicity, in which the unit cells are relatively poorly aligned, will produce broad reflections in the recorded diffraction images. The distance between reflections is inversely proportional to the unit cell axes lengths. Crystals with large unit cell axes therefore produce closely packed reflections and in order to distinguish between them the angle of rotation must be kept low. The X-ray dose a crystal receives for each diffraction image depends upon the strength of the beam and the length of exposure. A larger X-ray dose will produce a better signal to noise ratio as reflections get stronger and can increase the higher resolution limit of the data collected, up to the limit capable from a specific crystal. Unfortunately a large X-ray dose also has two unfavourable results reducing the completeness of the data. An increase in the amount of radiation damage suffered by the crystal leads to a loss of high resolution data over time, and reflections becoming too strong and overloading the detector can cause problems particularly with the stronger low resolution data. The X-ray dose must therefore be balanced between obtaining high resolution and completeness of the data.

14.8 Data processing

Once diffraction images have been collected the data must be indexed and integrated together to produce a complete dataset. Indexing is a process that estimates the dimensions and spacegroup of the unit cell by using the distances between reflections and the symmetry of the diffraction pattern. The solvent content of the crystal and number of molecules in the asymmetric unit can then be predicted based on the molecular weight of the protein and the unit cell dimensions and space group [197] (equation 14.2).

$$V_m = \frac{V}{M_r Z N} \qquad \text{(equation 14.2)}$$

The Matthews coefficient of a crystal, V_m, is calculated by dividing the volume of the unit cell, V, by the molecular weight of the protein, M_r, multiplied by the number of equivalent positions in the space group, Z , and the number of molecules in the ASU, N. The V_m can be calculated for different numbers of proteins in the ASU and compared to a database containing values derived from other protein crystals to provide an estimate of the most likely number of protein molecules in the ASU. Once a complete dataset has been collected and the

unit cell parameters have been determined the data is integrated together to produce a file containing all the data found during the experiment. The dataset can then be scaled to account for discrepancies in the data caused by external influences and equivalent reflections are merged. Following scaling the quality of the data can be assessed using a number of measures, most importantly, the completeness of the data, the signal to noise ratio (I/σI), the merging R-factor (R_{merge}) (equation 14.3) and the precision-indicating merging R-factor (R_{pim}) (equation 14.4).

$$R_{merge} = \frac{\sum_{hkl} \sum_{j=1}^{n} \left| I_{hkl.j} - \langle I_{hkl} \rangle \right|}{\sum_{hkl} \sum_{j=1}^{n} I_{hkl.j}} \quad \text{(equation 14.3)}$$

$$R_{pim} = \frac{\sum_{hkl} \sqrt{\frac{1}{n-1}} \sum_{j=1}^{n} \left| I_{hkl.j} - \langle I_{hkl} \rangle \right|}{\sum_{hkl} \sum_{j=1}^{n} I_{hkl.j}} \quad \text{(equation 14.4)}$$

R_{merge} and R_{pim} both provide a measure of the agreement between symmetry related reflections in a dataset. They differ in the application of a correction factor in R_{pim} that takes the multiplicity of the data into account. As would be expected the greater the number of measurements of a symmetry related reflection, the greater the discrepancy between them when summed together despite the increase in accuracy that comes from increased data.

14.9 Diffraction data to electron density

Each reflection in a diffraction image contains contributions from every atom within the unit cell and can be described by a structure factor (equation 14.5).

$$F_{hkl} = \sum_{j=1}^{n} f_j e^{2\pi i(hx_j + ky_j + lz_j)} \quad \text{(equation 14.5)}$$

The structure factor, F_{hkl}, is a Fourier sum of the contributions of each individual atom, j, to the reflection with indices, h, k, l, in the reciprocal lattice. The term, f_j, provides the scattering factor of the atom, j, treated as a sphere of electron density, giving the amplitude of its diffracted wave contribution, within the Fourier sum, dependent on the number of electrons present in the atom. The exponential component represents a simple three-

dimensional periodic function with both sine and cosine components. It relates to the position of atoms within the unit cell as x_j, y_j, z_j are the real space co-ordinates of atom, j, within the unit cell expressed as fractions of axis length. Alternatively F_{hkl} can be described as the contributions from volume elements, centred on co-ordinates x, y, z, (equation 14.6), or more simply as part of the unit cell (equation 14.7), of the electron density, ρ, within the unit cell.

$$F_{hkl} = \int_x \int_y \int_z \rho(x,y,z) e^{2\pi i(hx+ky+lz)} dx\, dy\, dz \qquad \text{(equation 14.6)}$$

$$F_{hkl} = \int_V \rho(x,y,z) e^{2\pi i(hx+ky+lz)} dv \qquad \text{(equation 14.7)}$$

The values of these elements are equal to the average electron density within the specified volume, so by decreasing the size of each individual volume element a value more exactly represents the true electron density at that point. As each term in a Fourier sum would represent an average electron density, the Fourier sum is instead described in terms of integrals in order to make the volume elements infinitely small and the corresponding electron density values reflect reality over the whole unit cell. Fourier transforms represent a reversible operation and the electron density can therefore be derived from a set of structure factors (equation 14.8).

$$\rho(x,y,z) = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl} e^{-2\pi i(hx+ky+lz)} \qquad \text{(equation 14.8)}$$

As the hkl values represent discrete entities, the equation contains a triple sum instead of the triple integral required for the continuous data of the previous equation (equation 14.7). Using this equation, an electron density map can be constructed provided the required information is available. Structure factors are periodic functions that possess an amplitude, frequency and phase and in order to calculate each structure factor, these three properties must be known. The amplitude is proportional to the square root of the intensity of each reflection, I_{hkl}, and the frequencies are provided by the values of h, k, l, and are therefore recorded in the diffraction image data. Unfortunately the phase information is lost during data collection and must be calculated using other methods.

14.10 Obtaining phases

A structure factor can be thought of in terms of a complex vector which can be split into amplitude, $|F_{hkl}|$, and phase, α_{hkl}, components. This allows the expression of the electron density in terms of structure factors (equation 14.9) where the phases have separate explicit values to the amplitudes (equation 14.9).

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F_{hkl}|e^{-2\pi i(hx+ky+lz)+i\alpha_{hkl}} \qquad \text{(equation 14.9)}$$

Three methods were used to calculate phases in the work described in this thesis. Isomorphous replacement and anomalous dispersion are experimental techniques that involve the determination of a substructure of certain marker atoms. The experimental amplitudes and calculated phases of the substructure are then combined with the experimental amplitudes of the protein diffraction data to calculate phases for the protein structure. Molecular replacement represents a theoretical technique that relies on the availability of a previously solved structure for a highly similar protein. The known structure is positioned in the unit cell and the phases are then calculated from the model. All three of these methods rely on the use of the Patterson function in some form.

14.10.1 The Patterson function

The Patterson function (equation 14.10) is a Fourier sum without phases using amplitudes that relate directly to reflection intensities in diffraction data. The function can therefore be used for diffraction data, despite the lack of phase information attached to each reflection.

$$P(u,v,w) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F_{hkl}|^2 e^{-2\pi i(hu+kv+lw)} \qquad \text{(equation 14.10)}$$

Each structure factor amplitude is squared, $|F_{hkl}|^2$, making it proportional to the measured intensities in diffraction data. The equation produces a Patterson map with its own co-ordinate system, u, v, w, representing vectors between atoms in the real space structure. If the structure of the unit cell is relatively simple, consisting of a small number of atoms, the Patterson function can therefore be used to determine the location of individual atoms in real

space based on the presence of vectors between them. However as the vectors found by Patterson methods are of an unknown direction, there are two possible solutions to the structure of atoms located using Patterson methods, typically referred to as the original and inverted hands. The atomic model produced can be used to calculate phases for the structure factors of the model (equation 14.5). Unfortunately the amount of information relating to vectors between atoms in a protein is huge as the number of atoms in even a relatively small protein of 20 kDa is in excess of 2,500 atoms meaning Patterson methods cannot be directly applied to entire protein structures.

14.10.2 Single isomorphous replacement and multiple isomorphous replacement

Isomorphous replacement represents one technique to effectively simplify the results of the Patterson function by reducing the number of atoms for which vectors are produced. The technique involves producing isomorphous crystals, identical to the native crystals, but for the addition of a small number of heavy atoms. This is achieved through co-crystallisation in the presence of heavy atoms, or through soaking crystals in a solution similar to the crystallisation condition but with added heavy atoms prior to mounting onto the data collection apparatus. Heavy atoms used in protein crystallography typically include iodine, mercury, gold or platinum which can bind to specific residues in the proteins structure. The native and derivative data can then be compared and if the native data is subtracted from the isomorphous derivative data an estimated diffraction pattern of the added heavy atom(s) within the unit cell is effectively produced (equation 14.11).

$$\mathrm{F}_H = \mathrm{F}_{PH} - \mathrm{F}_P \qquad \text{(equation 14.11)}$$

This process is combined with the Patterson function (equation 14.10) to create the difference Patterson function (equation 14.12).

$$\Delta P(u,v,w) = \frac{1}{V}\sum_h\sum_k\sum_l \Delta F_{hkl}^2 e^{-2\pi i(hu+kv+lw)} \qquad \text{(equation 14.12)}$$

The difference in the intensities of each reflection, ΔF, effectively ($|F_{PH}| - |F_P|$), are used to create a Patterson map for the heavy atom component, $\mathrm{F_H}$, of the derivative data. The resulting simple Patterson map can be used to solve a model of the positions of the heavy

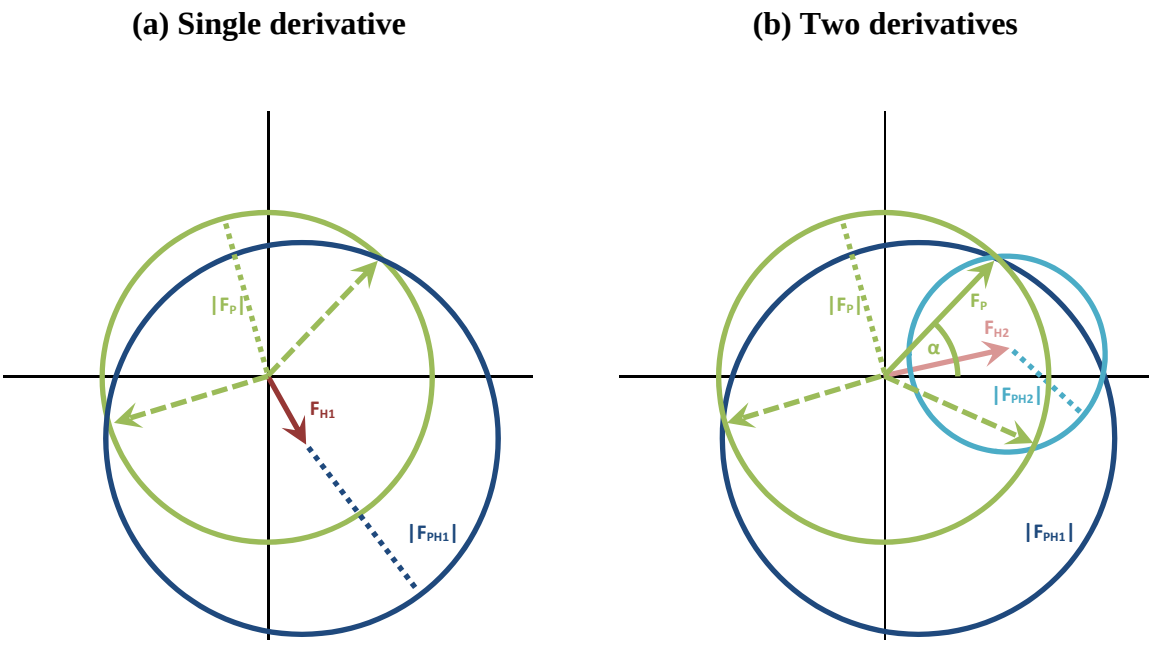


Figure 14.4 A graphical representation of phase calculation using isomorphous replacement. The green dotted lines, dashed arrows, arrow, circles and angle α represent $|F_P|$. The dark and light red arrows represent $|F_H|$ with both amplitude and calculated phases for derivatives one and two respectively. The dark and light blue dotted lines and circles represent $|F_{PH}|$ for derivatives one and two respectively. Solid arrows represent structure factors with known amplitude and phase; dashed arrows represent structure factors with known amplitude but of two possible phases; circles and dotted lines represent structure factors of known amplitude and unknown phase; and the angle α represents the desired protein phase information. **a** shows a vector diagram representing a solution for a single isomorphous derivative. The structure factors for the protein data, $|F_P|$, have known amplitude (dotted green line) but unknown phase. They can therefore be represented as a circle, centred on the origin, when plotted on axes representing the real and imaginary components of the structure factor (green circle). The heavy atom structure factor, F_{H1}, has a known amplitude and phase and is placed at the origin of the two axes (dark red arrow). A circle can be drawn around the end of the F_{H1} vector representing $|F_{PH1}|$ (dark blue circle) as these structure factors have a known amplitude (dark blue dotted line) but unknown phase. This provides two possible values for the phase of the protein structure factor, where the two circles intersect. **b** shows the same vector diagram including a second isomorphous derivative. Phases for the protein structure factor are derived using the same method, this time the heavy atom structure factor, F_{H2}, (light red arrow) is placed at the origin, and the derivative amplitudes $|F_{PH2}|$are plotted around its end (light blue circle). The location where the three circles intersect provides the correct phase, α, for the protein structure factor, F_P.

atom(s) within the unit cell. This model, complete with experimental amplitudes and calculated phases, coupled with the native and derivative experimental amplitudes from the diffraction data, can then be used to calculate phases for the protein structure. If the heavy atom structure factors, F_H, are subtracted from the derivative structure factors, F_{PH}, the resulting structure factors belong to the protein, F_P (equation 14.13).

$$F_P = F_{PH} - F_H \qquad \text{(equation 14.13)}$$

The phases of the protein structure can therefore be calculated, either diagrammatically using Harker constructs (figure 14.4) or computationally using X-ray crystallography software. A single derivative data set provides two possible estimates of the phase, α, for each structure factor, F_P (figure 14.4 a). To reach a single isomorphous replacement solution the value of the protein phase is taken as the centroid between the two possible values to provide a rough map of electron density that may be improvable by certain density modification techniques. Alternatively the phase ambiguity can be removed by using the multiple isomorphous replacement technique providing much better initial protein phases. Datasets are collected from different crystals with heavy atoms bound in different positions within the unit cell. In the simplest form a second derivative can be used to calculate a further two estimates for the protein phase. One of which should correspond to one of the phase estimates from the first derivative solving the ambiguity and providing a reasonable estimate for the phase of the reflection (figure 14.4 b).

14.10.3 Anomalous dispersion

The incorporation of atoms capable of anomalous dispersion in the unit cell allows for a further method to determine phases for the experimental data. Anomalous scattering occurs when the incident X-ray photon is absorbed by atoms in the crystal, causing a transition between electron orbitals, and reemitted with a different amplitude and phase as the energy is lost from the atom. The absorption of incident X-rays by an element is wavelength dependent with sharp rises at absorption edges equivalent to the energy required to effect specific electronic transitions (figure 14.5). Therefore overall scattering from the crystal has both a wavelength independent, but scattering angle dependent, contribution of every atom to diffraction, f_0, but also a wavelength dependent, but scattering angle independent, contribution from the anomalously scattering atoms, in the unit cell, f' and f'' (equation 14.14).

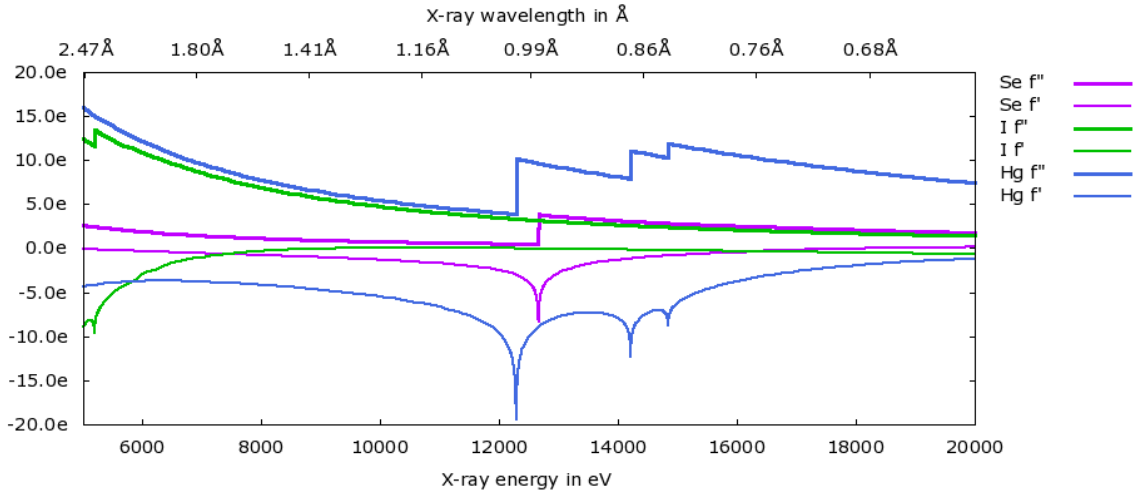


Figure 14.5 X-ray absorption edges of commonly used phasing elements. The figure shows the theoretical f' and f'' values for a single atom of an element at the given X-ray wavelength or energy for selenium, iodine and mercury. The figure was produced using the SSRL Absorption Package.

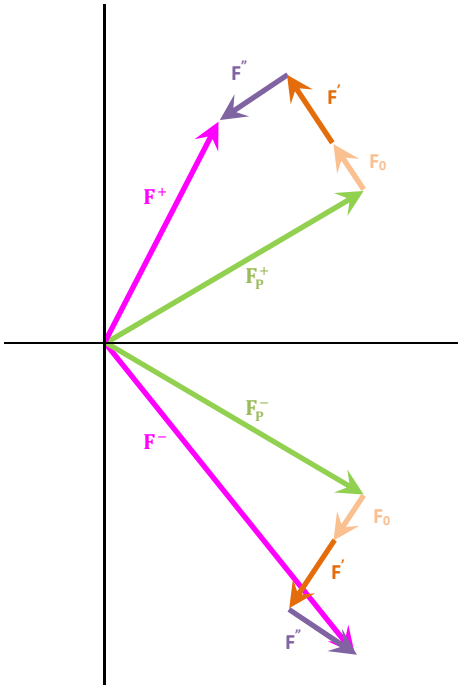


Figure 14.6 Anomalous scattering causes Friedel's law to breakdown. The green arrows represent the scattering from the non-anomalously scattering atoms (the protein) and are equal in amplitude but opposite in phase. The light orange and dark orange arrows represent the normal and dispersive scattering from the anomalously scattering atoms respectively. The purple arrows represent the phase shifted anomalous component of the scattering from the anomalously scattering atoms and it is this component that causes Friedel's law to breakdown. The pink arrows are the structure factors for the combined scattering of the non-anomalously and anomalously scattering atoms and it is the amplitudes of these that are directly measured in the diffraction data.

$$f_{anomalous} = f_0 + f' + if'' \qquad \text{(equation 14.14)}$$

The real dispersive difference component, f', alters the amplitude of the normal scattering while the imaginary anomalous difference component, f'', has a phase that is shifted by 90º to the normal, f_0, and dispersive, f', scattering. The effect of the anomalous f'' component of the scattering means that Friedel's rule is broken and Friedel pairs no longer have equal amplitudes and opposite phases becoming Bijovet pairs with values F^+ and F^- (figure 14.6).

14.10.4 Single-wavelength anomalous dispersion

In single-wavelength anomalous dispersion experiments the anomalously scattering atom substructure can be determined using Patterson methods. A difference Patterson function (equation 14.2) can be conducted using the amplitudes of the two members of the Bijovet pair, where ΔF, is effectively $(|F_{PA}^+| - |F_{PA}^-|)$, to locate the positions of the anomalously scattering atoms. The substructure model can be used to generate structure factors with an amplitude and phase for the normal scattering component from anomalously scattering atoms, $\mathbf{F}_A$. This can then be combined with the phase shifted anomalous component of the scattering from the anomalously scattering atoms and the two amplitudes of the members of a Bijovet pair to calculate two possible phases for the other non-anomalously scattering atoms in the crystal for each reflection (equation 14.15).

$$\mathbf{F}_{PA} = \mathbf{F}_P + \mathbf{F}_A + i\mathbf{F}_A'' \qquad \text{(equation 14.15)}$$

The total scattering of the protein and anomalously scattering atoms, $\mathbf{F}_{PA}$, is the sum of the protein scattering, $\mathbf{F}_P$, with the collinear normal, f_0, and real dispersive, f', elements from the anomalously scattering atoms, $\mathbf{F}_A$, and the imaginary anomalous component phase shifted by 90º, $i\mathbf{F}_A''$. Two possible protein phases can therefore be calculated in a similar fashion to that used in isomorphous replacement, either graphically (figure 14.7) or computationally using X-ray crystallography software.

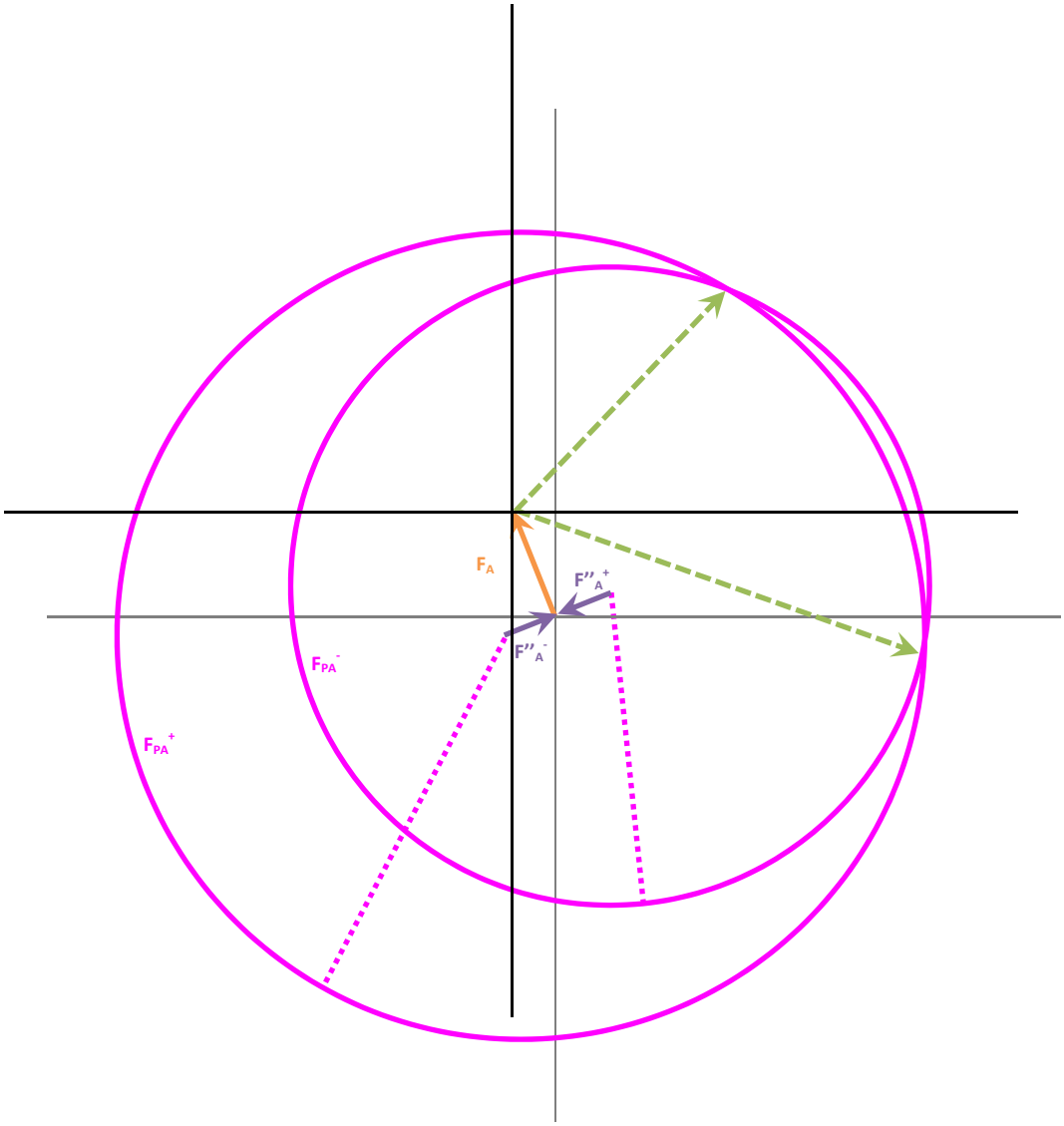


Figure 14.7 A graphical representation of phase calculation using single anomalous dispersion. This diagram is analogous to that for the isomorphous replacement solutions. The green dashed arrows represent possible values of $|F_P|$. The orange arrow represents $|F_A|$ with both amplitude and calculated phase. The pink dotted lines and circles represent $|F_{PA}^+|$ and $|F_{PA}^-|$. The vector diagram represents a solution for a SAD dataset. The normal scattering structure factors for the anomalously scattering atoms, F_A, has a known amplitude and phase and is placed at the origin of the two axes (orange arrow). The anomalous scattering contributions are also known, $F''_A{}^+$ and $F''_A{}^-$, are also known and plotted towards the origin. The structure factors for the combined anomalous and non-anomalously scattering atoms, $|F_{PA}^+|$ and $|F_{PA}^-|$, have known amplitude (dotted pink lines) but unknown phase. They can therefore be represented as a circle, centred on the end of the corresponding anomalous vector (pink circles). This provides two possible values for the phase of the non-anomalously scattering atoms structure factor, where the two circles intersect, with the structure factor being measured from the end of the anomalously scattering atom structure factor to the circles intersection.

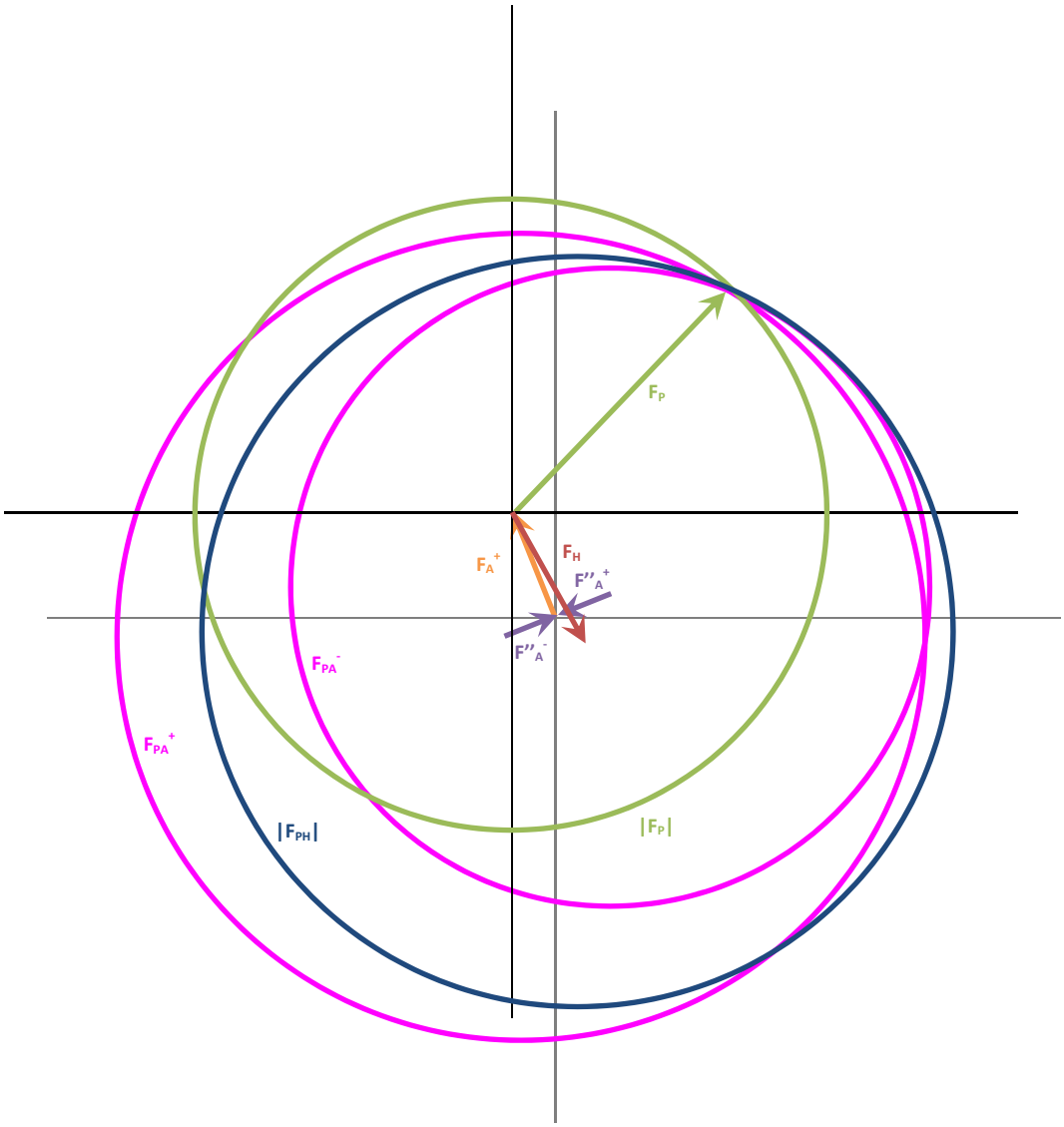


Figure 14.8 A graphical representation of phase calculation using single isomorphous replacement with anomalous scattering. This diagram is analogous to that for the SAD and isomorphous replacement solutions and is effectively constructed by overlaying figure 14.4 a with figure 14.7. This vector diagram provides a single value for the phase of the non-anomalously scattering atoms structure factor, where the three circles now intersect.

14.10.5 Single isomorphous replacement with anomalous scattering

SIR and SAD can be combined to form the single isomorphous replacement with anomalous scattering (SIRAS) technique in order to solve the phase ambiguity that arises from the two techniques. The possible protein phases found by SIR are symmetrical about the heavy atom phase, while protein phases found by SAD are symmetric perpendicular to the heavy atom phase. The two possible phases found using SIR and the two possible phases calculated by SAD should therefore have one solution in common (figure 14.8).

14.10.6 Multi-wavelength anomalous dispersion

The phase ambiguity from a SAD experiment can also be resolved by incorporating additional datasets collected at different wavelengths. This provides further information as different datasets will have different f' and f'' values, as these are wavelength dependent, but the same f_0, as this is wavelength independent. This information allows a single protein phase for each structure factor to be calculated (equations 14.16 and 14.7) [198].

$$|F^+|^2 = |F_T|^2 + a|F_A|^2 + b|F_T||F_A|\cos\Delta\varphi + c|F_T||F_A|\sin\Delta\varphi \quad \text{(equation 14.16)}$$

$$|F^-|^2 = |F_T|^2 + a|F_A|^2 + b|F_T||F_A|\cos\Delta\varphi - c|F_T||F_A|\sin\Delta\varphi \quad \text{(equation 14.17)}$$

The constants, a, b and c (equations 14.18, 14.19 and 14.20) are dependent on the scattering contributions from the anomalously scattering atoms.

$$a = \frac{f'^2 + f''^2}{{f_0}^2} \quad \text{(equation 14.18)}$$

$$b = \frac{2f'}{f_0} \quad \text{(equation 14.19)}$$

$$c = \frac{2f''}{f_0} \quad \text{(equation 14.20)}$$

The phase angle that is derived, $\Delta\varphi$, is not the phase angle of the protein, α or φ_T, but the difference between the phase of the scattering from the non-anomalously scattering atoms, φ_T, and the anomalously scattering atoms, φ_A. Providing data is available for multiple wavelengths containing an anomalous diffraction contribution there are sufficient data to directly derive protein phases. In multi-wavelength anomalous dispersion experiments, three wavelengths are typically collected and are referred to as the peak, inflection and remote datasets. The peak and inflection are close to the absorption edge and are collected to maximise f'' and minimise f' respectively with the remote dataset distant to the edge in order to maximise Δf' with respect to the inflection.

14.10.7 Molecular replacement

The final technique used in attempts to provide initial phases for structure solution in this thesis is molecular replacement. In order for molecular replacement to be successful a structure must already have been determined for a protein with a similar three dimensional fold. The known structure is used as a search model in order to obtain the phase information

for the target structure. Structure factors with both an amplitude and phase can be calculated for the search model. The phases from the model are then used as the initial estimate of the phases for the structure factors of the target protein alongside the experimentally determined amplitudes allowing an electron density map to be produced. The first step in solving a structure by molecular replacement is to create a suitable search model. This is usually achieved by finding a structure in the protein data bank with a high degree of sequence similarity before editing the model to better resemble the target protein. The second step requires the search model to be correctly positioned inside the unit cell of the target protein in order to calculate useful phases. Placing the model is further divided into two steps, the first is a search based on the orientation of the molecule(s) in the unit cell followed by a second search based on the location of the correctly orientated molecule(s) inside the unit cell. The rotation search is achieved through the use of Patterson functions which can be calculated for both the experimental amplitudes and the amplitudes of the known structure factors of the search model. The Patterson map from the experimental data is superimposed with maps created for the search model in a number of orientations, differing by small increments in the rotation of the molecule within a unit cell (equation 14.21).

$$R(\phi,\varphi,\chi) = \int_{u,v,w} P^t(u,v,w)P^m\{(u,v,w)(\phi,\varphi,\chi)\}\,du\,dv\,dw \qquad \text{(equation 14.21)}$$

The orientation of the search model, ϕ,φ,χ, that provides the greatest consensus between peaks in the experimental, P^t, and model, P^m, Patterson maps, $R(\phi,\varphi,\chi)$, is then taken for the translation search. This is performed by comparing the experimental amplitudes directly with calculated amplitudes from the correctly orientated model in the unit cell placed in a number of locations differing by small increments in their translations about the unit cell. One method of is to use the translation correlation function (equation 14.22).

$$C(t) = \frac{\sum(|\mathbf{F}_o|^2 - |\overline{\mathbf{F}_o}|^2)\left(|\mathbf{F}_c(t)|^2 - \left|\overline{\mathbf{F}_c(t)}\right|^2\right)}{\sqrt{(|\mathbf{F}_o|^2 - |\overline{\mathbf{F}_o}|^2)^2\left(|\mathbf{F}_c(t)|^2 - \left|\overline{\mathbf{F}_c(t)}\right|^2\right)^2}} \qquad \text{(equation 14.22)}$$

Structure factors are calculated for the protein, in the orientation found from the rotation search, in all positions, t, inside the cell on a small spaced grid. The correlation, C, between

the observed, $\mathbf{F}_o$, and calculated, $\mathbf{F}_c$, structure factors is then compared to find the best possible location within the unit cell. Finally the location that provides the highest degree of consensus is taken and the predicted phases of the model are combined with the experimental amplitudes to produce initial estimate for the protein structure factors.

14.10.8 Density modification

Initial calculated phases can be improved by applying prior knowledge about the density found in protein crystals and known or predicted properties specific to the crystal being studied. There are four main techniques that can be used for density modification procedures, solvent flattening and flipping, histogram matching and averaging the density around non-crystallographic symmetry elements.

14.11 Data processing, estimating phases and initial model production

Once a dataset was collected it was integrated using either XDS [199] or Mosflm [146] before scaling using the programmes XSCALE [199] or SCALA [188]. Phasing was attempted using a number of programmes and techniques for the work included in this thesis. Details of the techniques and programs used for each target are described in the relevant results sections. Experimental phasing by SIR, SAD, SIRAS and MAD were conducted using either the SHELX package of programmes (SHELXC, SHELXD and SHELXE with autotracing), the AutoSol programme of the Phenix suite [158] or the crank pipeline which uses programmes (AFRO/CRUNCH, BP3, SOLOMON, PARROT and Buccaneer) from the CCP4 suite of programmes [156]. Molecular replacement was conducted using Phaser [151] from the CCP4 suite [156]. Initial model building was done by the Buccaneer programme [187] of the CCP4 suite [156], or the Autobuild programme of the phenix suite of programmes [158].

14.12 Model rebuilding, refinement and validation

After initial phases for the protein have been calculated, and a starting protein model produced, the model is improved by iterative rounds of rebuilding and refinement. The aim of this process is to build the model that best describes the experimental data. The agreement between the model and the diffraction data is calculated by comparison of the observed structure factor amplitudes, from the experimental data, with a calculated set of structure factors amplitudes created from model to produce an R-factor, R (equation 14.23).

$$R = \frac{\sum_{hkl}\left||F_{obs}| - k|F_{calc}|\right|}{\sum_{hkl}|F_{obs}|} \quad \text{(equation 14.23)}$$

The calculated structure factors are adjusted by a scaling factor, k, to relate them to the same scale as the experimental data. A low R-factor is desirable meaning the model corresponds well with the data, with a value of zero meaning the model agrees perfectly with the data. In reality, due to errors in data collection and the limits of resolution, it is not possible to accurately determine the exact position of all the atoms in the unit cell. A free R-factor, R_{free}, is also calculated using a small random subset of the diffraction data that is excluded from the refinement process in order to monitor and guard against any incorporation of bias by overfitting. The R-factor and free R-factor should be a close to each other as possible as a difference of more than 5 % suggests the model has been over fitted to the data and become biased. Known structural parameters, based on theoretical and observed values from other high resolution structures, can also be used to gauge the quality of a model. These include chemical properties, such as bond lengths and angles, stereochemical properties, such as chirality, and conformational properties, such as atom clashes. Model improvement is achieved by cycles of adjusting the model to best fit an electron density map, followed by refinement and production of a new electron density map, which is in turn used to adjust the model. The progress while building a structure is monitored by a comparison of the R and R_{free} values from subsequent steps alongside analysis of the models geometry. The process of refinement makes small changes to the model, adjusting the positioning of atoms and their associated B-factors in order to reduce the overall R-factor. Refinement can be conducted with varying degrees of external restraints placed upon the model. The application of these restraints is decided based on the resolution of the data and the progress of model building. Generally, the lower the resolution of the data and the earlier in the model building process, the greater the number of these restraints that are applied.

14.13 Producing the final model

Refinement and rebuilding of models in this thesis were conducted using REFMAC5 [152], from the CCP4 programme suite [156], and COOT [159] respectively. The initial phased model was first examined in COOT and rebuilt to best fit the current electron density map while maintaining appropriate geometry between the atoms in the model. Rebuilding includes making alterations to existing components, and adding atoms corresponding to unbuilt parts of the protein and molecules from the surrounding buffer for which there is definitive density

in the map. Once the model could not obviously be further improved it was input alongside the original data into REFMAC5 [152] for refinement. Each refinement stage consisted of 10 cycles of maximum likelihood refinement provided with various external restraints. Two maps and an adjusted model were output at the end of each refinement stage. The model contains the minor adjustments made as part of the refinement process. The two maps are a weighted map of $2m|\mathbf{F}_{obs}| - D|\mathbf{F}_{calc}|$, where the weights are dependent on the quality of the current model, and $|\mathbf{F}_{obs}| - |\mathbf{F}_{calc}|$. The two maps were used in the COOT programme alongside the refined output model to guide further modification of the model before further refinement in an iterative process that finished when the model could not be improved further. During model building, the structural parameters were examined using the structure analysis tools within the COOT. Once a final structure had been completed, the geometry was analysed using PROCHECK [160] and the Molprobity server [161]. Molprobity also calculates a score based on the quality of the model geometry and compares this to other models in the PDB of a similar resolution in order to gauge the relative quality of the model.

15.0 Abbreviations and symbols

15.1 Crystallographic

Å	Angstrom (10^{-10} m)
a, b, c, α, β, γ	Real space unit cell dimensions and angles
ASU	Asymmetric unit
d_{hkl}	Inter-planar spacing in the reciprocal lattice
$\mathbf{F}$	Structure factor
$\mathbf{F}'$	Structure factor arising from the dispersive scattering from an atom
$\mathbf{F}''$	Structure factor arising from the anomalous scattering from an atom
$\mathbf{F}_0$	Structure factor arising from the normal scattering from an atom
$\mathbf{F}_{hkl}$	Structure factor for the reflection with indices hkl
$\mathbf{F}_A$	Structure factor arising from the anomalous contribution of all atoms
$\mathbf{F}_H$	Structure factor arising from the atoms in a heavy atom substructure
$\mathbf{F}_P$	Structure factor arising from the atoms in a protein structure
$\mathbf{F}_{PH}$	Structure factor arising from a protein structure containing heavy atoms
$\mathbf{F}_T$	Structure factor arising from the total contribution of all atoms
$\lvert\mathbf{F}^+\rvert, \lvert\mathbf{F}^-\rvert$	Structure factor amplitudes for reflections in a Bijovet pair
$\lvert\mathbf{F}_{hkl}\rvert$	Structure factor amplitude for the reflection with indices hkl
$\lvert\mathbf{F}_{hkl}\rvert$	Structure factor for the reflection with indices hkl
$\lvert\mathbf{F}_{calcs}\rvert$	Calculated structure factor amplitude
$\lvert\mathbf{F}_{obs}\rvert$	Observed structure factor amplitude
I_{hkl}	Intensity of the reflection with indices hkl
$I/\sigma I$	Signal to noise ratio

f'	Dispersive form factor from an atom
f''	Anomalous form factor from an atom
f_0	Normal form factor from an atom
$f_{anomalous}$	Atomic scattering from an anomalously scattering atom
hkl	Miller indices
MAD	Multiple-wavelength anomalous dispersion
MIR	Multiple isomorphous replacement
MR	Molecular replacement
R	R-factor
R_{free}	Free R-factor
R_{merge}	Merging R-factor
R_{pim}	Precision-indicating merging R-factor
SAD	Single-wavelength anomalous dispersion
SIR	Single isomorphous replacement
SIRAS	Single isomorphous replacement with anomalous scattering
u, v, w	co-ordinates in Patterson space
V_m	Matthews coefficient
x, y, z	co-ordinates in real space
Z	Number of equivalent positions in the unit cell
λ	Wavelength
φ_A	The phase of the structure factors from the anomalously scattering atoms
φ_T	The phase of the structure factors from all atoms
α	Phase angle of the structure factors from the protein molecule.

15.2 Biological and chemical

AHL	N-acyl-homoserine lactones
CPS	Capsular polysaccharide structure
DEAE	Diethylaminoethyl
DMSO	Dimethyl sulphoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide
EDTA	Ethylenediaminetetraacetic acid
EMTS	Sodium ethylmercuric thiosalicylate
HMAQ	4-hydroxy-3-methyl-2-alkylquinolones
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LPS	Lipopolysaccharide
MES	2-(N-morpholino)ethanesulfonic acid
mRNA	Messenger ribonucleic acid
NADPH	Nicotinamide adenine dinucleotide phosphate
PEG	polyethylene glycol
RNase	RNA nuclease
SDS	Sodium dodecyl sulfate
T3SS	Type 3 secretion system
T6SS	Type 6 secretion system
TRIS	TRIS(hydroxymethyl)aminomethane
Trx	Thioredoxin
X-gal	5-bromo-4-chloro-indolyl-β-D-galactopyranoside

15.3 Miscellaneous

AU	Absorbance unit
bp	Base pair (of nucleic acid)
CCD	Charge coupled device
K_{av}	Partition coefficient in gel filtration
LB media	Lysogeny broth media: 1 % (w/v) tryptone, 0.5 % (w/v) yeast extract, 1 % (w/v) NaCl
OD	Optical density
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PDB	Protein data bank
pI	Isoelectric point
RMSD	Root mean square deviation
SOC media	Super optimal brother with catabolite repression media: 2% (w/v) bacto-tryptone, 0.5% (w/v) Yeast extract, 10 mM NaCl, 2.5mM KCl
TAE buffer	TRIS acetate EDTA buffer

16.0 References

1. Whitmore, A. and C.S. Krishnaswami, *An account of the discovery of a hitherto undescribed infective disease occurring among the population of Rangoon.* Indian Med Gaz, 1912. **47**: p. 262-267.
2. Whitmore, A., *An account of a glanders-like disease occurring in Rangoon.* J. Hyg., 1913. **13**: p. 1-34.
3. Yabuuchi, E., Y. Kosako, H. Oyaizu, I. Yano, H. Hotta, Y. Hashimoto, T. Ezaki, and M. Arakawa, *Proposal of Burkholderia gen. nov. and transfer of seven species of the genus Pseudomonas homology group II to the new genus, with the type species Burkholderia cepacia (Palleroni and Holmes 1981) comb. nov.* Microbiology and Immunology, 1992. **36**(12): p. 1251-1275.
4. Paganin, P., S. Tabacchioni, and L. Chiarini, *Pathogenicity and biotechnological applications of the genus Burkholderia.* Central European Journal of Biology, 2011. **6**(6): p. 997-1005.
5. Ussery, D.W., K. Kiil, K. Lagesen, T. Sicheritz-Pontén, J. Bohlin, and T.M. Wassenaar, *The genus Burkholderia: Analysis of 56 genomic sequences*, H. Reuse and S. Bereswill, Editors. 2009. p. 140-157.
6. Zhu, B., S. Zhou, M. Lou, J. Zhu, B. Li, G. Xie, G. Jin, and R. de Mot, *Characterization and inference of gene gain/loss along burkholderia evolutionary history.* Evolutionary Bioinformatics, 2011. **2011**(7): p. 191-200.
7. Ho, C.C., C.C.Y. Lau, P. Martelli, S.Y. Chan, C.W.S. Tse, A.K.L. Wu, K.Y. Yuen, S.K.P. Lau, and P.C.Y. Woo, *Novel pan-genomic analysis approach in target selection for multiplex PCR identification and detection of Burkholderia pseudomallei, Burkholderia thailandensis, and Burkholderia cepacia complex species: A proof-of-concept study.* Journal of Clinical Microbiology, 2011. **49**(3): p. 814-821.
8. Schönmann, S., A. Loy, C. Wimmersberger, J. Sobek, C. Aquino, P. Vandamme, B. Frey, H. Rehrauer, and L. Eberl, *16S rRNA gene-based phylogenetic microarray for simultaneous identification of members of the genus Burkholderia.* Environmental Microbiology, 2009. **11**(4): p. 779-800.
9. Payne, G.W., P. Vandamme, S.H. Morgan, J.J. LiPuma, T. Coenye, A.J. Weightman, T.H. Jones, and E. Mahenthiralingam, *Development of a recA gene-based*

identification approach for the entire Burkholderia genus. Applied and Environmental Microbiology, 2005. **71**(7): p. 3917-3927.

10. Brett, P.J., D. DeShazer, and D.E. Woods, *Burkholderia thailandensis sp. nov., a Burkholderia pseudomallei-like species.* International Journal of Systematic Bacteriology, 1998. **48**(1): p. 317-320.
11. Pearson, T., P. Giffard, S. Beckstrom-Sternberg, R. Auerbach, H. Hornstra, A. Tuanyok, E.P. Price, M.B. Glass, B. Leadem, J.S. Beckstrom-Sternberg, G.J. Allan, J.T. Foster, D.M. Wagner, R.T. Okinaka, S.H. Sim, O. Pearson, Z. Wu, J. Chang, R. Kaul, A.R. Hoffmaster, T.S. Brettin, R.A. Robison, M. Mayo, J.E. Gee, P. Tan, B.J. Currie, and P. Keim, *Phylogeographic reconstruction of a bacterial species with high levels of lateral gene transfer.* BMC Biology, 2009. **7**.
12. Liguori, A.P., S.D. Warrington, J.L. Ginther, T. Pearson, J. Bowers, M.B. Glass, M. Mayo, V. Wuthiekanun, D. Engelthaler, S.J. Peacock, B.J. Currie, D.M. Wagner, P. Keim, and A. Tuanyok, *Diversity of 16s-23s rDNA internal transcribed spacer (its) reveals phylogenetic relationships in burkholderia pseudomallei and its near-neighbors.* PLoS ONE, 2011. **6**(12).
13. Choy, J.L., M. Mayo, A. Janmaat, and B.J. Currie, *Animal melioidosis in Australia.* Acta Tropica, 2000. **74**(2-3): p. 153-158.
14. Lee, Y.H., Y. Chen, X. Ouyang, and Y.H. Gan, *Identification of tomato plant as a novel host model for Burkholderia pseudomallei.* BMC Microbiology, 2010. **10**.
15. Cheng, A.C. and B.J. Currie, *Melioidosis: Epidemiology, pathophysiology, and management.* Clinical Microbiology Reviews, 2005. **18**(2): p. 383-416.
16. Chen, Y.S., S.C. Chen, C.M. Kao, and Y.L. Chen, *Effects of soil pH, temperature and water content on the growth of Burkholderia pseudomallei.* Folia Microbiologica, 2003. **48**(2): p. 253-256.
17. Dejsirilert, S., E. Kondo, D. Chiewsilp, and K. Kanai, *Growth and survival of Pseudomonas pseudomallei in acidic environments.* Japanese Journal of Medical Science and Biology, 1991. **44**(2): p. 63-74.
18. Wuthiekanun, V., M.D. Smith, and N.J. White, *Survival of Burkholderia pseudomallei in the absence of nutrients.* Transactions of the Royal Society of Tropical Medicine and Hygiene, 1995. **89**(5): p. 491.
19. Pumpuang, A., N. Chantratita, C. Wikraiphat, N. Saiprom, N.P.J. Day, S.J. Peacock, and V. Wuthiekanun, *Survival of Burkholderia pseudomallei in distilled water for 16*

years. Transactions of the Royal Society of Tropical Medicine and Hygiene, 2011. **105**(10): p. 598-600.

20. Gal, D., M. Mayo, H. Smith-Vaughan, P. Dasari, M. McKinnon, S.P. Jacups, A.I. Urquhart, M. Hassell, and B.J. Currie, *Contamination of hand wash detergent linked to occupationally acquired melioidosis.* American Journal of Tropical Medicine and Hygiene, 2004. **71**(3): p. 360-362.
21. Sookpranee, M., P. Lumbiganon, S. Puapermpoonsiri, A. Tattawasatra, and J. Nopwinyoovongs, *Contamination of Savlon solution with Pseudomonas pseudomallei at Srinagarind Hospital.* Melioidosis, 1989: p. 211-213.
22. Howard, K. and T.J.J. Inglis, *The effect of free chlorine on Burkholderia pseudomallei in potable water.* Water Research, 2003. **37**(18): p. 4425-4432.
23. Rose, L.J., E.W. Rice, B. Jensen, R. Murga, A. Peterson, R.M. Donlan, and M.J. Arduino, *Chlorine inactivation of bacterial bioterrorism agents.* Applied and Environmental Microbiology, 2005. **71**(1): p. 566-568.
24. Chantratita, N., V. Wuthiekanun, K. Boonbumrung, R. Tiyawisutsri, M. Vesaratchavest, D. Limmathurotsakul, W. Chierakul, S. Wongratanacheewin, S. Pukritiyakamee, N.J. White, N.P.J. Day, and S.J. Peacock, *Biological relevance of colony morphology and phenotypic switching by Burkholderia pseudomallei.* Journal of Bacteriology, 2007. **189**(3): p. 807-817.
25. Velapatiño, B., D. Limmathurotsakul, S.J. Peacock, and D.P. Speert, *Identification of differentially expressed proteins from Burkholderia pseudomallei isolated during primary and relapsing melioidosis.* Microbes and Infection, 2012. **14**(4): p. 335-340.
26. Chantratita, N., S. Tandhavanant, C. Wikraiphat, L.A. Trunck, D.A. Rholl, A. Thanwisai, N. Saiprom, D. Limmathurotsakul, S. Korbsrisate, N.P.J. Day, H.P. Schweizer, and S.J. Peacock, *Proteomic analysis of colony morphology variants of Burkholderia pseudomallei defines a role for the arginine deiminase system in bacterial survival.* Journal of Proteomics, 2012. **75**(3): p. 1031-1042.
27. Dance, D.A.B., *Melioidosis as an emerging global problem.* Acta Tropica, 2000. **74**(2-3): p. 115-119.
28. Inglis, T.J.J., D.B. Rolim, and A.D.Q. Sousa, *Melioidosis in the Americas.* American Journal of Tropical Medicine and Hygiene, 2006. **75**(5): p. 947-954.
29. Currie, B.J., *Advances and remaining uncertainties in the epidemiology of Burkholderia pseudomallei and melioidosis.* Transactions of the Royal Society of Tropical Medicine and Hygiene, 2008. **102**(3): p. 225-227.

30. Wiersinga, W.J., T. van der Poll, N.L. White, N.P. Day, and S.J. Peacock, *Melioidosis: Insights into the pathogenicity of Burkholderia pseudomallei.* Nature Reviews Microbiology, 2006. **4**(4): p. 272-282.
31. Stevens, M.P., J.M. Stevens, R.L. Jeng, L.A. Taylor, M.W. Wood, P. Hawes, P. Monaghan, M.D. Welch, and E.E. Galyov, *Identification of a bacterial factor required for actin-based motility of Burkholderia pseudomallei.* Molecular Microbiology, 2005. **56**(1): p. 40-53.
32. Sun, G.W., J. Lu, S. Pervaiz, W.P. Cao, and Y.H. Gan, *Caspase-1 dependent macrophage death induced by Burkholderia pseudomallei.* Cellular Microbiology, 2005. **7**(10): p. 1447-1458.
33. Currie, B.J., M. Mayo, N.M. Anstey, P. Donohoe, A. Haase, and D.J. Kemp, *A cluster of melioidosis cases from an endemic region is clonal and is linked to the water supply using molecular typing of Burkholderia pseudomallei isolates.* American Journal of Tropical Medicine and Hygiene, 2001. **65**(3): p. 177-179.
34. Abbink, F.C., J.M. Orendi, and A.J. De Beaufort, *Mother-to-child transmission of Burkholderia pseudomallei.* New England Journal of Medicine, 2001. **344**(15): p. 1171-1172.
35. McCormick, J.B., D.J. Sexton, and J.G. McMurray, *Human to human transmission of Pseudomonas pseudomallei.* Annals of Internal Medicine, 1975. **83**(4): p. 512-513.
36. Kanaphun, P., N. Thirawattanasuk, Y. Suputtamongkol, P. Naigowit, D.A.B. Dance, M.D. Smith, and N.J. White, *Serology and carriage of Pseudomonas pseudomallei: A prospective study in 1000 hospitalized children in Northeast Thailand.* Journal of Infectious Diseases, 1993. **167**(1): p. 230-233.
37. Yee, K.C., M.K. Lee, C.T. Chua, and S.D. Puthucheary, *Melioidosis, the great mimicker: A report of 10 cases from Malaysia.* Journal of Tropical Medicine and Hygiene, 1988. **91**(5): p. 249-254.
38. White, N.J., W. Chaowagul, V. Wuthiekanun, D.A.B. Dance, Y. Wattanagoon, and N. Pitakwatchara, *Halving of mortality of severe melioidosis by ceftazidime.* Lancet, 1989. **2**(8665): p. 697-701.
39. White, N.J., *Melioidosis.* Lancet, 2003. **361**(9370): p. 1715-1722.
40. Behera, B., T.L.V.D. Prasad Babu, A. Kamalesh, and G. Reddy, *Ceftazidime resistance in Burkholderia pseudomallei: First report from India.* Asian Pacific Journal of Tropical Medicine, 2012. **5**(4): p. 329-330.

41. Chantratita, N., D.A. Rholl, B. Sim, V. Wuthiekanun, D. Limmathurotsakul, P. Amornchai, A. Thanwisai, H.H. Chua, W.F. Ooi, M.T.G. Holden, N.P. Day, P. Tan, H.P. Schweizer, and S.J. Peacock, *Antimicrobial resistance to ceftazidime involving loss of penicillin-binding protein 3 in Burkholderia pseudomallei.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(41): p. 17165-17170.

42. Sarovich, D.S., E.P. Price, A.T. von Schulze, J.M. Cook, M. Mayo, L.M. Watson, L. Richardson, M.L. Seymour, A. Tuanyok, D.M. Engelthaler, T. Pearson, S.J. Peacock, B.J. Currie, P. Keim, and D.M. Wagner, *Characterization of ceftazidime resistance mechanisms in clinical isolates of burkholderia pseudomallei from Australia.* PLoS ONE, 2012. **7**(2).

43. Gatedee, J., K. Kritsiriwuthinan, E.E. Galyov, J. Shan, E. Dubinina, N. Intarak, M.R. Clokie, and S. Korbsrisate, *Isolation and characterization of a novel podovirus which infects burkholderia pseudomallei.* Virology Journal, 2011. **8**.

44. Sarkar-Tyson, M. and R.W. Titball, *Progress toward development of vaccines against melioidosis: A review.* Clinical Therapeutics, 2010. **32**(8): p. 1437-1445.

45. Peacock, S.J., D. Limmathurotsakul, Y. Lubell, G.C.K.W. Koh, L.J. White, N.P.J. Day, and R.W. Titball, *Melioidosis vaccines: A systematic review and appraisal of the potential to exploit biodefense vaccines for public health purposes.* PLoS Neglected Tropical Diseases, 2012. **6**(1).

46. Nieves, W., S. Asakrah, O. Qazi, K.A. Brown, J. Kurtz, D.P. Aucoin, J.B. McLachlan, C.J. Roy, and L.A. Morici, *A naturally derived outer-membrane vesicle vaccine protects against lethal pulmonary Burkholderia pseudomallei infection.* Vaccine, 2011. **29**(46): p. 8381-8389.

47. AuCoin, D.P., D.E. Reed, N.L. Marlenee, R.A. Bowen, P. Thorkildson, B.M. Judy, A.G. Torres, and T.R. Kozel, *Polysaccharide specific monoclonal antibodies provide passive protection against intranasal challenge with burkholderia pseudomallei.* PLoS ONE, 2012. **7**(4).

48. Rotz, L.D., A.S. Khan, S.R. Lillibridge, S.M. Ostroff, and J.M. Hughes, *Public health assessment of potential biological terrorism agents.* Emerging Infectious Diseases, 2002. **8**(2): p. 225-230.

49. Holden, M.T.G., R.W. Titball, S.J. Peacock, A.M. Cerdeño-Tárraga, T. Atkins, L.C. Crossman, T. Pitt, C. Churcher, K. Mungall, S.D. Bentley, M. Sebaihia, N.R. Thomson, M. Bason, I.R. Beacham, K. Brooks, K.A. Brown, N.F. Brown, G.L.

Challis, I. Cherevach, T. Chillingworth, A. Cronin, B. Crossett, P. Davis, D. DeShazer, T. Feltwell, A. Fraser, Z. Hance, H. Hauser, S. Holroyd, K. Jagels, K.E. Keith, M. Maddison, S. Moule, C. Price, M.A. Quail, E. Rabbinowitsch, K. Rutherford, M. Sanders, M. Simmonds, S. Songsivilai, K. Stevens, S. Tumapa, M. Vesaratchavest, S. Whitehead, C. Yeats, B.G. Barrell, P.C.F. Oyston, and J. Parkhill, *Genomic plasticity of the causative agent of melioidosis, Burkholderia pseudomallei.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(39): p. 14240-14245.

50. Gamage, A.M., G. Shui, M.R. Wenk, and K.L. Chua, *N-Octanoylhomoserine lactone signalling mediated by the BpsI-BpsR quorum sensing system plays a major role in biofilm formation of Burkholderia pseudomallei.* Microbiology, 2011. **157**(4): p. 1176-1186.

51. Ulrich, R.L., D. DeShazer, E.E. Brueggemann, H.B. Hines, P.C. Oyston, and J.A. Jeddeloh, *Role of quorum sensing in the pathogenicity of Burkholderia pseudomallei.* Journal of Medical Microbiology, 2004. **53**(11): p. 1053-1064.

52. Chan, Y.Y. and K.L. Chua, *The Burkholderia pseudomallei BpeAB-OprB efflux pump: Expression and impact on quorum sensing and virulence.* Journal of Bacteriology, 2005. **187**(14): p. 4707-4719.

53. Ying, Y.C., S.B. Hao, T.M.C. Tan, M.E. Mattmann, G.D. Geske, J. Igarashi, T. Hatano, H. Suga, H.E. Blackwell, and L.C. Kim, *Control of quorum sensing by a Burkholderia pseudomallei multidrug efflux pump.* Journal of Bacteriology, 2007. **189**(11): p. 4320-4324.

54. Vial, L., F. Lépine, S. Milot, M.C. Groleau, V. Dekimpe, D.E. Woods, and E. Déziel, *Burkholderia pseudomallei, B. thailandensis, and B. ambifaria produce 4-hydroxy-2-alkylquinoline analogues with a methyl group at the 3 position that is required for quorum-sensing regulation.* Journal of Bacteriology, 2008. **190**(15): p. 5339-5352.

55. Korbsrisate, S., M. Vanaporn, P. Kerdsuk, W. Kespichayawattana, P. Vattanaviboon, P. Kiatpapan, and G. Lertmemongkolchai, *The Burkholderia pseudomallei RpoE (AlgU) operon is involved in environmental stress tolerance and biofilm formation.* FEMS Microbiology Letters, 2005. **252**(2): p. 243-249.

56. Vanaporn, M., P. Vattanaviboon, V. Thongboonkerd, and S. Korbsrisate, *The rpoE operon regulates heat stress response in Burkholderia pseudomallei.* FEMS Microbiology Letters, 2008. **284**(2): p. 191-196.

57. Thongboonkerd, V., M. Vanaporn, N. Songtawee, R. Kanlaya, S. Sinchaikul, S.T. Chen, A. Easton, K. Chu, G.J. Bancroft, and S. Korbsrisate, *Altered proteome in Burkholderia pseudomallei rpoE operon knockout mutant: Insights into mechanisms of rpoE operon in stress tolerance, survival, and virulence.* Journal of Proteome Research, 2007. **6**(4): p. 1334-1341.

58. Subsin, B., M.S. Thomas, G. Katzenmeier, J.G. Shaw, S. Tungpradabkul, and M. Kunakorn, *Role of the Stationary Growth Phase Sigma Factor RpoS of Burkholderia pseudomallei in Response to Physiological Stress Conditions.* Journal of Bacteriology, 2003. **185**(23): p. 7008-7014.

59. Jangiam, W., S. Loprasert, D.R. Smith, and S. Tungpradabkul, *Burkholderia pseudomallei Rpos regulates OxyR and the katG-dpsA operon under conditions of oxidative stress.* Microbiology and Immunology, 2010. **54**(7): p. 389-397.

60. Utaisincharoen, P., S. Arjcharoen, K. Limposuwan, S. Tungpradabkul, and S. Sirisinha, *Burkholderia pseudomallei RpoS regulates multinucleated giant cell formation and inducible nitric oxide synthase expression in mouse macrophage cell line (RAW 264.7).* Microbial Pathogenesis, 2006. **40**(4): p. 184-189.

61. Lengwehasatit, I., A. Nuchtas, S. Tungpradabkul, S. Sirisinha, and P. Utaisincharoen, *Involvement of B. pseudomallei RpoS in apoptotic cell death in mouse macrophages.* Microbial Pathogenesis, 2008. **44**(3): p. 238-245.

62. Wongtrakoongate, P., S. Tumapa, and S. Tungpradabkul, *Regulation of a quorum sensing system by stationary phase sigma factor RpoS and their co-regulation of target genes in Burkholderia pseudomallei.* Microbiology and Immunology, 2012. **56**(5): p. 281-294.

63. Puthucheary, S.D., J. Vadivelu, C. Ce-Cile, W. Kum-Thong, and G. Ismail, *Short report: Electron microscopic demonstration of extracellular structure of Burkholderia pseudomallei.* American Journal of Tropical Medicine and Hygiene, 1996. **54**(3): p. 313-314.

64. Kawahara, K., S. Dejsirilert, and T. Ezaki, *Characterization of three capsular polysaccharides produced by Burkholderia pseudomallei.* FEMS Microbiology Letters, 1998. **169**(2): p. 283-287.

65. Nimtz, M., V. Wray, T. Domke, B. Brenneke, S. Häussler, and I. Steinmetz, *Structure of an acidic exopolysaccharide of Burkholderia pseudomallei.* European Journal of Biochemistry, 1997. **250**(2): p. 608-616.

66. Masoud, H., M. Ho, T. Schollaardt, and M.B. Perry, *Characterization of the capsular polysaccharide of Burkholderia (Pseudomonas) pseudomallei 304b.* Journal of Bacteriology, 1997. **179**(18): p. 5663-5669.
67. Perry, M.B., L.L. MacLean, T. Schollaardt, L.E. Bryan, and M. Ho, *Structural characterization of the lipopolysaccharide O antigens of Burkholderia pseudomallei.* Infection and Immunity, 1995. **63**(9): p. 3348-3352.
68. Knirel, Y.A., N.A. Paramonov, A.S. Shashkov, N.K. Kochetkov, R.G. Yarullin, S.M. Farber, and V.I. Efremenko, *Structure of the polysaccharide chains of Pseudomonas pseudomallei lipopolysaccharides.* Carbohydrate Research, 1992. **233**: p. 185-193.
69. Reckseidler-Zenteno, S.L., D.F. Viteri, R. Moore, E. Wong, A. Tuanyok, and D.E. Woods, *Characterization of the type III capsular polysaccharide produced by Burkholderia pseudomallei.* Journal of Medical Microbiology, 2010. **59**(12): p. 1403-1414.
70. Reckseidler-Zenteno, S.L., R. DeVinney, and D.E. Woods, *The capsular polysaccharide of Burkholderia pseudomallei contributes to survival in serum by reducing complement factor C3b deposition.* Infection and Immunity, 2005. **73**(2): p. 1106-1115.
71. Reckseidler, S.L., D. DeShazer, P.A. Sokol, and D.E. Woods, *Detection of bacterial virulence genes by subtractive hybridization: Identification of capsular polysaccharide of Burkholderia pseudomallei as a major virulence determinant.* Infection and Immunity, 2001. **69**(1): p. 34-44.
72. Sarkar-Tyson, M., J.E. Thwaite, S.V. Harding, S.J. Smither, P.C.F. Oyston, T.P. Atkins, and R.W. Titball, *Polysaccharides and virulence of Burkholderia pseudomallei.* Journal of Medical Microbiology, 2007. **56**(8): p. 1005-1010.
73. Kawahara, K., S. Dejsirilert, H. Danbara, and T. Ezaki, *Extraction and characterization of lipopolysaccharide from Pseudomonas pseudomallei.* FEMS Microbiology Letters, 1992. **96**(2-3): p. 129-134.
74. DeShazer, D., P.J. Brett, and D.E. Woods, *The type II O-antigenic polysaccharide moiety of Burkholderia pseudomallei lipopolysaccharide is required for serum resistance and virulence.* Molecular Microbiology, 1998. **30**(5): p. 1081-1100.
75. Moore, R.A., A. Tuanyok, and D.E. Woods, *Survival of Burkholderia pseudomallei in Water.* BMC Research Notes, 2008. **1**.

76. Vorachit, M., K. Lam, P. Jayanetra, and J.W. Costerton, *Electron microscopy study of the mode of growth of Pseudomonas pseudomallei in vitro and in vivo.* Journal of Tropical Medicine and Hygiene, 1995. **98**(6): p. 379-391.

77. Taweechaisupapong, S., C. Kaewpa, C. Arunyanart, P. Kanla, P. Homchampa, S. Sirisinha, T. Proungvitaya, and S. Wongratanacheewin, *Virulence of Burkholderia pseudomallei does not correlate with biofilm formation.* Microbial Pathogenesis, 2005. **39**(3): p. 77-85.

78. Sawasdidoln, C., S. Taweechaisupapong, R.W. Sermswan, U. Tattawasart, S. Tungpradabkul, and S. Wongratanacheewin, *Growing Burkholderia pseudomallei in biofilm stimulating conditions significantly induces antimicrobial resistance.* PLoS ONE, 2010. **5**(2).

79. Kumar, A., M. Mayo, L.A. Trunck, A.C. Cheng, B.J. Currie, and H.P. Schweizer, *Expression of resistance-nodulation-cell-division efflux pumps in commonly used Burkholderia pseudomallei strains and clinical isolates from northern Australia.* Transactions of the Royal Society of Tropical Medicine and Hygiene, 2008. **102**(SUPPL. 1): p. S145-S151.

80. Moore, R.A., D. Deshazer, S. Reckseidler, A. Weissman, and D.E. Woods, *Efflux-mediated aminoglycoside and macrolide resistance in Burkholderia pseudomallei.* Antimicrobial Agents and Chemotherapy, 1999. **43**(3): p. 465-470.

81. Trunck, L.A., K.L. Propst, V. Wuthiekanun, A. Tuanyok, S.M. Beckstrom-Sternberg, J.S. Beckstrom-Sternberg, S.J. Peacock, P. Keim, S.W. Dow, and H.P. Schweizer, *Molecular basis of rare aminoglycoside susceptibility and pathogenesis of Burkholderia pseudomalleiclinical isolates from Thailand.* PLoS Neglected Tropical Diseases, 2009. **3**(9).

82. Mima, T. and H.P. Schweizer, *The BpeAB-OprB efflux pump of Burkholderia pseudomallei 1026b does not play a role in quorum sensing, virulence factor production, or extrusion of aminoglycosides but is a broad-spectrum drug efflux system.* Antimicrobial Agents and Chemotherapy, 2010. **54**(8): p. 3113-3120.

83. Kumar, A., K.L. Chua, and H.P. Schweizer, *Method for regulated expression of single-copy efflux pump genes in a surrogate Pseudomonas aeruginosa strain: Identification of the BpeEF-OprC chloramphenicol and trimethoprim efflux pump of Burkholderia pseudomallei 1026b.* Antimicrobial Agents and Chemotherapy, 2006. **50**(10): p. 3460-3463.

84. Fehlner-Gardiner, C.C. and M.A. Valvano, *Cloning and characterization of the Burkholderia vietnamiensis norM gene encoding a multi-drug efflux protein.* FEMS Microbiology Letters, 2002. **215**(2): p. 279-283.

85. Jones, A.L., T.J. Beveridge, and D.E. Woods, *Intracellular survival of Burkholderia pseudomallei.* Infection and Immunity, 1996. **64**(3): p. 782-790.

86. Mohamed, R., S. Nathan, N. Embi, N. Razak, and G. Ismail, *Inhibition of macromolecular synthesis in cultured macrophages by Pseudomonas pseudomallei exotoxin.* Microbiology and Immunology, 1989. **33**(10): p. 811-820.

87. Vanaporn, M., M. Wand, S.L. Michell, M. Sarkar-Tyson, P. Ireland, S. Goldman, C. Kewcharoenwong, D. Rinchai, G. Lertmemongkolchai, and R.W. Titball, *Superoxide dismutase C is required for intracellular survival and virulence of burkholderia pseudomallei.* Microbiology, 2011. **157**(8): p. 2392-2400.

88. Loprasert, S., W. Whangsuk, R. Sallabhan, and S. Mongkolsuk, *DpsA protects the human pathogen Burkholderia pseudomallei against organic hydroperoxide.* Archives of Microbiology, 2004. **182**(1): p. 96-101.

89. Loprasert, S., R. Sallabhan, W. Whangsuk, and S. Mongkolsuk, *Compensatory increase in ahpC gene expression and its role in protecting Burkholderia pseudomallei against reactive nitrogen intermediates.* Archives of Microbiology, 2003. **180**(6): p. 498-502.

90. Chuaygud, T., S. Tungpradabkul, S. Sirisinha, K.L. Chua, and P. Utaisincharoen, *A role of Burkholderia pseudomallei flagella as a virulent factor.* Transactions of the Royal Society of Tropical Medicine and Hygiene, 2008. **102**(SUPPL. 1): p. S140-S144.

91. Chua, K.L., Y.Y. Chan, and Y.H. Gan, *Flagella are virulence determinants of Burkholderia pseudomallei.* Infection and Immunity, 2003. **71**(4): p. 1622-1629.

92. Deshazer, D., P.J. Brett, R. Carlyon, and D.E. Woods, *Mutagenesis of Burkholderia pseudomallei with Tn5-OT182: Isolation of motility mutants and molecular characterization of the flagellin structural gene.* Journal of Bacteriology, 1997. **179**(7): p. 2116-2125.

93. Breitbach, K., K. Rottner, S. Klocke, M. Rohde, A. Jenzora, J. Wehland, and I. Steinmetz, *Actin-based motility of Burkholderia pseudomallei involves the Arp 2/3 complex, but not N-WASP and Ena/VASP proteins.* Cellular Microbiology, 2003. **5**(6): p. 385-393.

94. Kespichayawattana, W., S. Rattanachetkul, T. Wanun, P. Utaisincharoen, and S. Sirisinha, *Burkholderia pseudomallei induces cell fusion and actin-associated membrane protrusion: A possible mechanism for cell-to-cell spreading.* Infection and Immunity, 2000. **68**(9): p. 5377-5384.

95. Boddey, J.A., C.J. Day, C.P. Flegg, R.L. Ulrich, S.R. Stephens, I.R. Beacham, N.A. Morrison, and I.R.A. Peak, *The bacterial gene lfpA influences the potent induction of calcitonin receptor and osteoclast-related genes in Burkholderia pseudomallei-induced TRAP-positive multinucleated giant cells.* Cellular Microbiology, 2007. **9**(2): p. 514-531.

96. Harley, V.S., D.A.B. Dance, B.S. Drasar, and G. Tovey, *Effects of Burkholderia pseudomallei and other Burkholderia species on eukaryotic cells in tissue culture.* Microbios, 1998. **96**(384): p. 71-93.

97. Wong, K.T., S.D. Puthucheary, and J. Vadivelu, *The histopathology of human melioidosis.* Histopathology, 1995. **26**(1): p. 51-55.

98. Boddey, J.A., C.P. Flegg, C.J. Day, I.R. Beacham, and I.R. Peak, *Temperature-regulated microcolony formation by Burkholderia pseudomallei requires pilA and enhances association with cultured human cells.* Infection and Immunity, 2006. **74**(9): p. 5374-5381.

99. Essex-Lopresti, A.E., J.A. Boddey, R. Thomas, M.P. Smith, M.G. Hartley, T. Atkins, N.F. Brown, C.H. Tsang, I.R.A. Peak, J. Hill, I.R. Beacham, and R.W. Titball, *A type IV pilin, pilA, contributes to adherence of Burkholderia pseudomallei and virulence in vivo.* Infection and Immunity, 2005. **73**(2): p. 1260-1264.

100. Ahmed, K., H.D.R. Enciso, H. Masaki, M. Tao, A. Omori, P. Tharavichikul, and T. Nagatake, *Attachment of Burkholderia pseudomallei to pharyngeal epithelial cells: A highly pathogenic bacteria with low attachment ability.* American Journal of Tropical Medicine and Hygiene, 1999. **60**(1): p. 90-93.

101. Ashdown, L.R. and J.M. Koehler, *Production of hemolysin and other extracellular enzymes by clinical isolates of Pseudomonas pseudomallei.* Journal of Clinical Microbiology, 1990. **28**(10): p. 2331-2334.

102. Gauthier, Y.P., F.M. Thibault, J.C. Paucod, and D.R. Vidal, *Protease production by Burkholderia pseudomallei and virulence in mice.* Acta Tropica, 2000. **74**(2-3): p. 215-220.

103. Lee, M.A. and Y. Liu, *Sequencing and characterization of a novel serine metalloprotease from Burkholderia pseudomallei.* FEMS Microbiology Letters, 2000. **192**(1): p. 67-72.

104. Chin, C.Y., R. Othman, and S. Nathan, *The Burkholderia pseudomallei serine protease MprA is autoproteolytically activated to produce a highly stable enzyme.* Enzyme and Microbial Technology, 2007. **40**(2): p. 370-377.

105. Valade, E., F.M. Thibault, Y.P. Gauthier, M. Palencia, M.Y. Popoff, and D.R. Vidal, *The PmlI-PmlR Quorum-Sensing System in Burkholderia pseudomallei Plays a Key Role in Virulence and Modulates Production of the MprA Protease.* Journal of Bacteriology, 2004. **186**(8): p. 2288-2294.

106. Ling, J.M.L., S. Nathan, L.K. Hin, and R. Mohamed, *Purification and Characterisation of a Burkholderia pseudomallei Protease Expressed in Recombinant E. coli.* Journal of Biochemistry and Molecular Biology, 2001. **34**(6): p. 509-516.

107. Tumwasorn, S., K. Lertpocasombat, and K. Saithanu, *A purified protease from Pseudomonas pseudomallei produces dermonecrosis in guinea pigs.* Selected Papers from the First International Symposium on Melioidosis, 1994: p. 70-73.

108. Rainbow, L., M.C. Wilkinson, P.J. Sargent, C.A. Hart, and C. Winstanley, *Identification and Expression of a Burkholderia pseudomallei Collagenase in Escherichia coli.* Current Microbiology, 2004. **48**(4): p. 300-304.

109. Korbsrisate, S., A.P. Tomaras, S. Damnin, J. Ckumdee, V. Srinon, I. Lengwehasatit, M.L. Vasil, and S. Suparak, *Characterization of two distinct phospholipase C enzymes from Burkholderia pseudomallei.* Microbiology, 2007. **153**(6): p. 1907-1915.

110. Tuanyok, A., M. Tom, J. Dunbar, and D.E. Woods, *Genome-wide expression analysis of Burkholderia pseudomallei infection in a hamster model of acute melioidosis.* Infection and Immunity, 2006. **74**(10): p. 5465-5476.

111. Yang, H., W. Chaowagul, and P.A. Sokol, *Siderophore production by Pseudomonas pseudomallei.* Infection and Immunity, 1991. **59**(3): p. 776-780.

112. Alice, A.F., C.S. López, C.A. Lowe, M.A. Ledesma, and J.H. Crosa, *Genetic and transcriptional analysis of the siderophore malleobactin biosynthesis and transport genes in the human pathogen Burkholderia pseudomallei K96243.* Journal of Bacteriology, 2006. **188**(4): p. 1551-1566.

113. Yang, H., C.D. Kooi, and P.A. Sokol, *Ability of Pseudomonas pseudomallei malleobactin to acquire transferrin- bound, lactoferrin-bound, and cell-derived iron.* Infection and Immunity, 1993. **61**(2): p. 656-662.

114. Tuanyok, A., H.S. Kim, W.C. Nierman, Y. Yu, J. Dunbar, R.A. Moore, P. Baker, M. Tom, J.M.L. Ling, and D.E. Woods, *Genome-wide expression analysis of iron regulation in Burkholderia pseudomallei and Burkholderia mallei using DNA microarrays.* FEMS Microbiology Letters, 2005. **252**(2): p. 327-335.

115. DeShazer, D., P.J. Brett, M.N. Burtnick, and D.E. Woods, *Molecular characterization of genetic loci required for secretion of exoproducts in Burkholderia pseudomallei.* Journal of Bacteriology, 1999. **181**(15): p. 4661-4664.

116. Warawa, J. and D.E. Woods, *Type III secretion system cluster 3 is required for maximal virulence of Burkholderia pseudomallei in a hamster infection model.* FEMS Microbiology Letters, 2005. **242**(1): p. 101-108.

117. Stevens, M.P., M.W. Wood, L.A. Taylor, P. Monaghan, P. Hawes, P.W. Jones, T.S. Wallis, and E.E. Galyov, *An Inv/Mxi-Spa-like type III protein secretion system in Burkholderia pseudomallei modulates intracellular behaviour of the pathogen.* Molecular Microbiology, 2002. **46**(3): p. 649-659.

118. Suparak, S., W. Kespichayawattana, A. Haque, A. Easton, S. Damnin, G. Lertmemongkolchai, G.J. Bancroft, and S. Korbsrisate, *Multinucleated giant cell formation and apoptosis in infected host cells is mediated by Burkholderia pseudomallei type III secretion protein BipB.* Journal of Bacteriology, 2005. **187**(18): p. 6556-6560.

119. D'Cruze, T., L. Gong, P. Treerat, G. Ramm, J.D. Boyce, M. Prescott, B. Adler, and R.J. Devenish, *Role for the Burkholderia pseudomallei type three secretion system cluster 1 bpscn gene in virulence.* Infection and Immunity, 2011. **79**(9): p. 3659-3664.

120. Stevens, M.P., A. Haque, T. Atkins, J. Hill, M.W. Wood, A. Easton, M. Nelson, C. Underwood-Fowler, R.W. Titball, G.J. Bancroft, and E.E. Galyov, *Attenuated virulence and protective efficacy of a Burkholderia pseudomallei bsa type III secretion mutant in murine models of melioidosis.* Microbiology, 2004. **150**(8): p. 2669-2676.

121. Cullinane, M., L. Gong, X. Li, N. Lazar-Adler, T. Tra, E. Wolvetang, M. Prescott, J.D. Boyce, R.J. Devenish, and B. Adler, *Stimulation of autophagy suppresses the intracellular survival of Burkholderia pseudomallei in mammalian cell lines.* Autophagy, 2008. **4**(6): p. 744-753.

122. Stevens, M.P., A. Friebel, L.A. Taylor, M.W. Wood, P.J. Brown, W.D. Hardt, and E.E. Galyov, *A Burkholderia pseudomallei type III secreted protein, BopE, facilitates*

bacterial invasion of epithelial cells and exhibits guanine nucleotide exchange factor activity. Journal of Bacteriology, 2003. **185**(16): p. 4992-4996.

123. Muangman, S., S. Korbsrisate, V. Muangsombut, V. Srinon, N.L. Adler, G.N. Schroeder, G. Frankel, and E.E. Galyov, *BopC is a type III secreted effector protein of Burkholderia pseudomallei.* FEMS Microbiology Letters, 2011. **323**(1): p. 75-82.

124. Yao, Q., J. Cui, Y. Zhu, G. Wang, L. Hu, C. Long, R. Cao, X. Liu, N. Huang, S. Chen, L. Liu, and F. Shao, *A bacterial type III effector family uses the papain-like hydrolytic activity to arrest the host cell cycle.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(10): p. 3716-3721.

125. Shalom, G., J.G. Shaw, and M.S. Thomas, *In vivo expression technology identifies a type VI secretion system locus in Burkholderia pseudomallei that is induced upon invasion of macrophages.* Microbiology, 2007. **153**(8): p. 2689-2699.

126. Pilatz, S., K. Breitbach, N. Hein, B. Fehlhaber, J. Schulze, B. Brenneke, L. Eberl, and I. Steinmetz, *Identification of Burkholderia pseudomallei genes required for the intracellular life cycle and in vivo virulence.* Infection and Immunity, 2006. **74**(6): p. 3576-3586.

127. Burtnick, M.N., P.J. Brett, S.V. Harding, S.A. Ngugi, W.J. Ribot, N. Chantratita, A. Scorpio, T.S. Milne, R.E. Dean, D.L. Fritz, S.J. Peacock, J.L. Prior, T.P. Atkins, and D. DeShazer, *The cluster 1 type VI secretion system is a major virulence determinant in Burkholderia pseudomallei.* Infection and Immunity, 2011. **79**(4): p. 1512-1525.

128. Fisher, N.A., W.J. Ribot, W. Applefeld, and D. DeShazer, *The Madagascar hissing cockroach as a novel surrogate host for Burkholderia pseudomallei, B. mallei and B. thailandensis.* BMC Microbiology, 2012: p. 117.

129. Cruz-Migoni, A., G.M. Hautbergue, P.J. Artymiuk, P.J. Baker, M. Bokori-Brown, C.T. Chang, M.J. Dickman, A. Essex-Lopresti, S.V. Harding, N.M. Mahadi, L.E. Marshall, G.W. Mobbs, R. Mohamed, S. Nathan, S.A. Ngugi, C. Ong, W.F. Ooi, L.J. Partridge, H.L. Phillips, M.F. Raih, S. Ruzheinikov, M. Sarkar-Tyson, S.E. Sedelnikova, S.J. Smither, P. Tan, R.W. Titball, S.A. Wilson, and D.W. Rice, *A Burkholderia pseudomallei toxin inhibits helicase activity of translation factor eIF4A.* Science, 2011. **334**(6057): p. 821-824.

130. Wongtrakoongate, P., N. Mongkoldhumrongkul, S. Chaijan, S. Kamchonwongpaisan, and S. Tungpradabkul, *Comparative proteomic profiles and the potential markers between Burkholderia pseudomallei and Burkholderia thailandensis.* Molecular and Cellular Probes, 2007. **21**(2): p. 81-91.

131. Osiriphun, Y., P. Wongtrakoongate, S. Sanongkiet, P. Suriyaphol, V. Thongboonkerd, and S. Tungpradabkul, *Identification and characterization of RpoS regulon and RpoS-dependent promoters in Burkholderia pseudomallei.* Journal of Proteome Research, 2009. **8**(6): p. 3118-3131.

132. Wongtrakoongate, P., S. Roytrakul, S. Yasothornsrikul, and S. Tungpradabkul, *A proteome reference map of the causative agent of melioidosis Burkholderia pseudomallei.* Journal of Biomedicine and Biotechnology, 2011. **2011**.

133. Kim, Y., L. Volkart, J. Abdullah, and A. Joachimiak, *PDB ID: 2OEZ Crystal Structure of the Protein of Unknown Function VP2528 from Vibrio parahaemolyticus.*

134. Harding, S.V., M. Sarkar-Tyson, S.J. Smither, T.P. Atkins, P.C.F. Oyston, K.A. Brown, Y. Liu, R. Wait, and R.W. Titball, *The identification of surface proteins of Burkholderia pseudomallei.* Vaccine, 2007. **25**(14): p. 2664-2672.

135. Thompson, D.B., K. Crandall, S.V. Harding, S.J. Smither, G.B. Kitto, R.W. Titball, and K.A. Brown, *In silico analysis of potential diagnostic targets from Burkholderia pseudomallei.* Transactions of the Royal Society of Tropical Medicine and Hygiene, 2008. **102**(SUPPL. 1): p. S61-S65.

136. Sofia, H.J., G. Chen, B.G. Hetzler, J.F. Reyes-Spindola, and N.E. Miller, *Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods.* Nucleic Acids Research, 2001. **29**(5): p. 1097-1106.

137. Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins, *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-2948.

138. Falquet, L., L. Bordoli, V. Ioannidis, M. Pagni, and C.V. Jongeneel, *Swiss EMBnet node web server.* Nucleic Acids Research, 2003. **31**(13): p. 3782-3783.

139. Petersen, T.N., S. Brunak, G. Von Heijne, and H. Nielsen, *SignalP 4.0: Discriminating signal peptides from transmembrane regions.* Nature Methods, 2011. **8**(10): p. 785-786.

140. Gasteiger, E., C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, and A. Bairoch, *Protein identification and analysis tools on the ExPASy server*, in *The Proteomics Protocols Handbook*, J.M. Walker, Editor 2005, Humana Press. p. 571-607.

141. Kelley, L.A. and M.J.E. Sternberg, *Protein structure prediction on the Web: a case study using the Phyre server.* Nat. Protocols, 2009. **4**(3): p. 363-371.

142. Wimmerova, M., E. Mitchell, J.F. Sanchez, C. Gautier, and A. Imberty, *Crystal structure of fungal lectin. Six-bladed β-propeller fold and novel fucose recognition mode for Aleuria aurantia lectin.* Journal of Biological Chemistry, 2003. **278**(29): p. 27059-27067.

143. Cioci, G., E.P. Mitchell, V. Chazalet, H. Debray, S. Oscarson, M. Lahmann, C. Gautier, C. Breton, S. Perez, and A. Imberty, *β-Propeller crystal structure of Psathyrella velutina lectin: An integrin-like fungal protein interacting with monosaccharides and calcium.* Journal of Molecular Biology, 2006. **357**(5): p. 1575-1591.

144. Wilson, J.J., O. Matsushita, A. Okabe, and J. Sakon, *A bacterial collagen-binding domain with novel calcium-binding motif controls domain orientation.* EMBO Journal, 2003. **22**(8): p. 1743-1752.

145. Baskaran, N., R.P. Kandpal, A.K. Bhargava, M.W. Glynn, A. Bale, and S.M. Weissman, *Uniform amplification of a mixture of deoxyribonucleic acids with varying GC content.* Genome Research, 1996. **6**(7): p. 633-638.

146. Leslie, A.G.W., *Recent changes to the MOSFLM package for processing film and image plate data.* Recent Changes to the MOSFLM Package for Processing Film and Image Plate Data, 1992.

147. Winter, G., *xia2: an expert system for macromolecular crystallography data reduction.* Journal of Applied Crystallography, 2010. **43**(1).

148. Kabsch, W., *Evaluation of single-crystal X-ray diffraction data from a position-sensitive detector.* J. Appl. Crystallogr., 1988. **21**: p. 916-924.

149. Kantardjieff, K.A. and B. Rupp, *Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals.* Prot Science, 2003. **12**: p. 1865-1871.

150. Stein, N., *CHAINSAW: a program for mutating pdb files used as templates in molecular replacement.* Journal of Applied Crystallography, 2008. **41**(3): p. 641-643.

151. McCoy, A.J., R.W. Grosse-Kunstleve, P.D. Adams, M.D. Winn, L.C. Storoni, and R.J. Read, *Phaser crystallographic software.* Journal of Applied Crystallography, 2007. **40**(4): p. 658-674.

152. Murshudov, G.N., P. Skubak, A.A. Lebedev, N.S. Pannu, R.A. Steiner, R.A. Nicholls, M.D. Winn, F. Long, and A.A. Vagin, *REFMAC5 for the refinement of*

macromolecular crystal structures. Acta Crystallographica Section D, 2011. **67**(4): p. 355-367.

153. Evans, P., *Scaling and assessment of data quality.* Acta Crystallographica Section D: Biological Crystallography, 2006. **62**(1): p. 72-82.
154. Sheldrick, G., *Experimental phasing with SHELXC/D/E: combining chain tracing with density modification.* Acta Crystallographica Section D, 2010. **66**(4): p. 479-485.
155. Pape, T. and T. Schneider, *{HKL2MAP}: a graphical user interface for phasing with {SHELX} programs.* J. Appl. Cryst., 2004. **37**: p. 843-844.
156. Winn, M.D., C.C. Ballard, K.D. Cowtan, E.J. Dodson, P. Emsley, P.R. Evans, R.M. Keegan, E.B. Krissinel, A.G.W. Leslie, A. McCoy, S.J. McNicholas, G.N. Murshudov, N.S. Pannu, E.A. Potterton, H.R. Powell, R.J. Read, A. Vagin, and K.S. Wilson, *Overview of the CCP4 suite and current developments.* Acta Crystallographica Section D: Biological Crystallography, 2011. **67**(4): p. 235-242.
157. Evans, G. and R. Pettifer, *{CHOOCH}: a program for deriving anomalous-scattering factors from {X-ray} fluorescence spectra.* J. Appl. Cryst., 2001. **34**: p. 82-86.
158. Adams, P.D., P.V. Afonine, G. Bunkóczi, V.B. Chen, I.W. Davis, N. Echols, J.J. Headd, L.W. Hung, G.J. Kapral, R.W. Grosse-Kunstleve, A.J. McCoy, N.W. Moriarty, R. Oeffner, R.J. Read, D.C. Richardson, J.S. Richardson, T.C. Terwilliger, and P.H. Zwart, *PHENIX: A comprehensive Python-based system for macromolecular structure solution.* Acta Crystallographica Section D: Biological Crystallography, 2010. **66**(2): p. 213-221.
159. Emsley, P., B. Lohkamp, W.G. Scott, and K. Cowtan, *Features and development of Coot.* Acta Crystallographica Section D: Biological Crystallography, 2010. **66**(4): p. 486-501.
160. Laskowski, R.A., M.W. MacArthur, D.S. Moss, and J.M. Thornton, *PROCHECK: a program to check the stereochemical quality of protein structures.* Journal of Applied Crystallography, 1993. **26**(2): p. 283-291.
161. Chen, V.B., W.B. Arendall Iii, J.J. Headd, D.A. Keedy, R.M. Immormino, G.J. Kapral, L.W. Murray, J.S. Richardson, and D.C. Richardson, *MolProbity: All-atom structure validation for macromolecular crystallography.* Acta Crystallographica Section D: Biological Crystallography, 2010. **66**(1): p. 12-21.
162. Holm, L. and P. Rosenström, *Dali server: Conservation mapping in 3D.* Nucleic Acids Research, 2010. **38**(SUPPL. 2): p. W545-W549.

163. Krissinel, E. and K. Henrick, *Inference of Macromolecular Assemblies from Crystalline State.* Journal of Molecular Biology, 2007. **372**(3): p. 774-797.

164. Janin, J., *Specific versus non-specific contacts in protein crystals.* Nat Struct Mol Biol, 1997. **4**(12): p. 973-974.

165. Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, *Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.* Nucleic Acids Research, 1997. **25**(17): p. 3389-3402.

166. Sievers, F., A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, and D.G. Higgins, *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Molecular Systems Biology, 2011. **7**.

167. Buchanan, B.B., A. Holmgren, J.P. Jacquot, and R. Scheibe, *Fifty years in the thioredoxin field and a bountiful harvest.* Biochimica et Biophysica Acta - General Subjects, 2012. **1820**(11): p. 1822-1829.

168. Eisenreich, W., K. Kemter, A. Bacher, S.B. Mulrooney, C.H. Williams Jr, and F. MÃ¼ller, *13C-, 15N- and 31P-NMR studies of oxidized and reduced low molecular mass thioredoxin reductase and some mutant proteins.* European Journal of Biochemistry, 2004. **271**(8): p. 1437-1452.

169. Lennon, B.W. and C.H. Williams Jr, *Effect of pyridine nucleotide on the oxidative half-reaction of Escherichia coli thioredoxin reductase.* Biochemistry, 1995. **34**(11): p. 3670-3677.

170. Lennon, B.W. and C.H. Williams Jr, *Enzyme-monitored turnover of Escherichia coli thioredoxin reductase: Insights for catalysis.* Biochemistry, 1996. **35**(15): p. 4704-4712.

171. Mulrooney, S.B. and C.H. Williams Jr, *Potential active-site base of thioredoxin reductase from Escherichia coli: Examination of histidine245 and aspartate139 by site-directed mutagenesis.* Biochemistry, 1994. **33**(11): p. 3148-3154.

172. O'Donnell, M.E. and C.H. Williams Jr, *Proton stoichiometry in the reduction of the FAD and disulfide of Escherichia coli thioredoxin reductase. Evidence for a base at the active site.* Journal of Biological Chemistry, 1983. **258**(22): p. 13795-13805.

173. O'Donnell, M.E. and C.H. Williams Jr, *Reconstitution of Escherichia coli thioredoxin reductase with 1-deazaFAD. Evidence for 1-deazaFAD C-4a adduct formation linked to the ionization of an active site base.* Journal of Biological Chemistry, 1984. **259**(4): p. 2243-2251.

174. O'Donnell, M.E. and C.H. Williams Jr, *Reaction of both active site thiols of reduced thioredoxin reductase with N-ethylmaleimide.* Biochemistry, 1985. **24**(26): p. 7617-7621.

175. Prongay, A.J. and C.H. Williams Jr, *Evidence for direct interaction between cysteine 138 and the flavin in thioredoxin reductase. A study using flavin analogs.* Journal of Biological Chemistry, 1990. **265**(31): p. 18968-18975.

176. Zanetti, G. and C.H. Williams Jr, *Characterization of the active center of thioredoxin reductase.* Journal of Biological Chemistry, 1967. **242**(22): p. 5232-5236.

177. Lennon, B.W., C.H. Williams Jr, and M.L. Ludwig, *Crystal structure of reduced thioredoxin reductase from Escherichia coli: Structural flexibility in the isoalloxazine ring of the flavin adenine dinucleotide cofactor.* Protein Science, 1999. **8**(11): p. 2366-2379.

178. Waksman, G., T.S.R. Krishna, C.H. Williams Jr, and J. Kuriyan, *Crystal structure of Escherichia coli thioredoxin reductase refined at 2 Å resolution. Implications for a large conformational change during catalysis.* Journal of Molecular Biology, 1994. **236**(3): p. 800-816.

179. Kuriyan, J., T.S.R. Krishna, L. Wong, B. Guenther, A. Pahler, C.H. Williams Jr, and P. Model, *Convergent evolution of similar function in two structurally divergent enzymes.* Nature, 1991. **352**(6331): p. 172-174.

180. Lennon, B.W. and C.H. Williams Jr, *Reductive half-reaction of thioredoxin reductase from Escherichia coli.* Biochemistry, 1997. **36**(31): p. 9464-9477.

181. Wang, P.F., D.M. Veine, S.H. Ahn, and C.H. Williams Jr, *A stable mixed disulfide between thioredoxin reductase and its substrate, thioredoxin: Preparation and characterization.* Biochemistry, 1996. **35**(15): p. 4812-4819.

182. Mulrooney, S.B. and C.H. Williams Jr, *Evidence for two conformational states of thioredoxin reductase from Escherichia coli: Use of intrinsic and extrinsic quenchers of flavin fluorescence as probes to observe domain rotation.* Protein Science, 1997. **6**(10): p. 2188-2195.

183. Veine, D.M., S.B. Mulrooney, P.F. Wang, and C.H. Williams Jr, *Formation and properties of mixed disulfides between thioredoxin reductase from Escherichia coli and thioredoxin: Evidence that cysteine-138 functions to initiate dithiol-disulfide interchange and to accept the reducing equivalent from reduced flavin.* Protein Science, 1998. **7**(6): p. 1441-1450.

184. Veine, D.M., K. Ohnishi, and C.H. Williams Jr, *Thioredoxin reductase from Escherichia coli: Evidence of restriction to a single conformation upon formation of a crosslink between engineered cysteines.* Protein Science, 1998. **7**(2): p. 369-375.

185. Lennon, B.W., Williams C.H, Jr., and M.L. Ludwig, *Twists in catalysis: Alternating conformations of Escherichia coli thioredoxin reductase.* Science, 2000. **289**(5482): p. 1190-1194.

186. Corsini, L., M. Hothorn, G. Stier, V. Rybin, K. Scheffzek, T.J. Gibson, and M. Sattler, *Dimerization and protein binding specificity of the U2AF homology motif of the splicing factor Puf60.* Journal of Biological Chemistry, 2009. **284**(1): p. 630-639.

187. Cowtan, K., *The Buccaneer software for automated model building. 1. Tracing protein chains.* Acta Crystallographica Section D, 2006. **62**(9): p. 1002-1011.

188. Diederichs, K. and P.A. Karplus, *Improved R-factors for diffraction data analysis in macromolecular crystallography.* Nature Structural Biology, 1997. **4**(4): p. 269-275.

189. Katti, S.K., D.M. LeMaster, and H. Eklund, *Crystal structure of thioredoxin from Escherichia coli at 1.68 Å resolution.* Journal of Molecular Biology, 1990. **212**(1): p. 167-184.

190. Collet, J.F. and J. Messens, *Structure, function, and mechanism of thioredoxin proteins.* Antioxidants and Redox Signaling, 2010. **13**(8): p. 1205-1216.

191. Chandonia, J.-M. and S.E. Brenner, *The Impact of Structural Genomics: Expectations and Outcomes.* Science, 2006. **311**(5759): p. 347-351.

192. Slabinski, L., L. Jaroszewski, L. Rychlewski, I.A. Wilson, S.A. Lesley, and A. Godzik, *XtalPred: a web server for prediction of protein crystallizability.* Bioinformatics, 2007. **23**(24): p. 3403-5.

193. Slabinski, L., L. Jaroszewski, A.P.C. Rodrigues, L. Rychlewski, I.A. Wilson, S.A. Lesley, and A. Godzik, *The challenge of protein structure determination--lessons from structural genomics.* Protein Sci, 2007. **16**(11): p. 2472-82.

194. Kibbe, W.A., *OligoCalc: an online oligonucleotide properties calculator.* Nucleic Acids Research, 2007. **35**(suppl 2): p. W43-W46.

195. Bikandi, J., R.S. Millán, A. Rementeria, and J. Garaizar, *In silico analysis of complete bacterial genomes: PCR, AFLP–PCR and endonuclease restriction.* Bioinformatics, 2004. **20**(5): p. 798-799.

196. Vincze, T., J. Posfai, and R.J. Roberts, *NEBcutter: a program to cleave DNA with restriction enzymes.* Nucleic Acids Research, 2003. **31**(13): p. 3688-3691.

197. Matthews, B.W., *Solvent content of protein crystals.* J Mol Biol, 1968. **33**: p. 491-497.

198. Hendrickson, W.A., J.L. Smith, and S. Sheriff, *Direct phase determination based on anomalous scattering*, in *Methods in Enzymology*, C.H.W.H.S.N.T. Harold W. Wyckoff, Editor 1985, Academic Press. p. 41-55.
199. Kabsch, W., *Integration, scaling, space-group assignment and post-refinement.* Acta Crystallographica Section D: Biological Crystallography, 2010. **66**(2): p. 133-144.